中国研究生创新实践系列大赛
“华为杯”第二十届中国研究生数学建模竞赛

学　　校	清华大学，清华大学深圳国际研究生院
参赛队号	**23900310014**
队员姓名	**1.** 范柯雨
	2. 陈炳杰
	3. 张婉婷

中国研究生创新实践系列大赛
“华为杯”第二十届中国研究生数学建模竞赛

题　目： 出血性脑卒中临床智能诊疗建模

摘　要：

出血性脑卒中是一种严重的神经疾病，在其发展过程中，血肿扩张和血肿周围的水肿都是预后不良的重要危险因素。本文的研究基于 160 名血性脑卒中患者的临床数据和多次影像检查结果，构建了血肿扩张、水肿体积进展以及患者预后等预测模型，并探索了不同特征因素与预后预测的关联关系。

对于**问题一**，我们对**血肿扩张概率预测**进行了建模与分析。具体来说：

在**子问题 (a)** 中，基于患者发病和多次影像检查时间，判断了前 100 位患者是否在发病后 48 小时内发生了**血肿扩张事件**，并记录了血肿扩张事件的发生时间。

在**子问题 (b)** 中，以患者的个人史、疾病史和发病相关信息为输入，经过**数据标准化、缺失值处理、Spearman 相关性分析、PCA** 等处理后，构建预测模型来预测患者发生血肿扩张的概率。我们比较了 **XGBoost、随机森林和 Logistic 回归**三种模型，最终选择准确率高达 **90%** 的 XGBoost 模型，预测了所有患者发生血肿扩张的概率。

对于**问题二**，我们对**水肿体积随时间进展**进行了建模与分析，并研究了个体差异和不同治疗方法的影响。具体来说：

在**子问题 (a)** 中，关注了前 100 个患者水肿体积随时间的变化，我们使用了 **RBF 核函数**和**样条插值**方法来拟合全体患者的**水肿体积随时间进展曲线**，同时计算了拟合**残差**，分析结果说明 RBF 核函数拟合的效果更佳。

在**子问题 (b)** 中，提出患者水肿体积变化线条-线条三维趋势点-Mean-Shift 聚类的算法，捕捉个体差异，并据此分成**四个不同的进展趋势亚组**。使用 RBF 核函数方法来拟合这些亚组的水肿体积曲线，计算得到相较于子问题 (a) 降低了 **34.3%** 的平均残差效果。

在**子问题 (c)** 中，使用**决策树分析**和**统计学分析**两种方法，探索了不同**治疗方法**对水肿体积进展模式的影响。

在**子问题 (d)** 中，探索了血肿体积、水肿体积和治疗方法之间的关系，通过相关性分析以及统计图表分析得到了**不同因素之间的相互作用**。

对于**问题三**，我们基于患者的个人史、疾病史、治疗方法和影像特征，对**患者预后预测**进行了建模与分析，并探索了与各特征之间的关联关系。具体来说：

在**子问题 (a)** 中，基于**首次**影像结果建立出血性脑卒中患者 90 天 mRS 评分的预测模型。我们使用了**独热编码**和数据拼接来处理多源数据，通过对比**随机森林**和 **XGBoost** 两种机器学习模型，得出 XGBoost 算法的预测效果更佳。

在**子问题 (b)** 中，基于**所有**影像结果建立出血性脑卒中患者 90 天 mRS 评分的预测模型。我们使用了 **XGBoost** 和 **RNN** 模型进行更全面的预测，模型预测准确性相对于问题 (a) 有了大幅提升，并且 RNN 模型对患者的 90 天 mRS 评分预测效果高达 **98%**。

在**子问题 (c)** 中，使用 Spearman 系数进行**相关性分析**，探讨了患者的个人史、疾病史、治疗方法以及多个影像特征与 90 天 mRS 评分之间的关联关系。通过识别与患者预后相关的关键因素，为临床相关决策提供建议。

本研究的结果强调了临床特征和影像检查的重要性，以及机器学习方法在出血性脑卒中患者管理中的潜在应用。我们的研究为改善出血性脑卒中患者的预后评估提供了有力支持，有望在未来的临床实践中发挥积极作用。

关键词： 出血性脑卒中　患者预后　机器学习　XGBoost　预测模型

目录

1 问题背景与问题重述 5

1.1 问题背景 5

1.2 数据介绍 5

1.3 问题重述 8

2 模型假设与符号说明 10

2.1 模型假设 10

2.2 符号说明 10

3 问题一：血肿扩张预测建模及分析 11

3.1 问题分析 11

3.2 子问题一：48 小时血肿扩张判断 12

3.3 子问题二：血肿扩张概率预测 12

3.3.1 数据预处理 12

3.3.2 主成分分析 14

3.3.3 模型构建 15

3.3.4 模型训练 18

3.3.5 模型对比 18

4 问题二：水肿体积进展建模及分析 19

4.1 问题分析 19

4.2 子问题一：全体患者水肿体积进展曲线 19

4.2.1 水肿体积随时间变化分布 19

4.2.2 水肿体积随时间进展曲线拟合 20

4.3 子问题二：不同人群的水肿体积进展曲线 23

4.4 子问题三：不同治疗方法对水肿体积进展的影响 27

4.4.1 决策树分析 27

4.4.2 统计学分析 29

4.4.3 总结 30

4.5 子问题四：血肿体积、水肿体积及治疗方法的关系 30

5 问题三：预后预测建模及分析 35

5.1 问题分析 35
5.2 子问题一：基于首次影像的 mRS 评分预测模型 35
5.2.1 数据预处理 35
5.2.2 模型构建 36
5.2.3 结果分析 36
5.3 子问题二：基于所有影像的 mRS 评分预测模型 36
5.3.1 XGBoost 模型 37
5.3.2 RNN 模型 37
5.3.3 模型对比 39
5.4 子问题三：分析预后与关键因素的关联 40
5.4.1 个人史和疾病史相关因素 40
5.4.2 治疗方法相关因素 40
5.4.3 影像特征相关因素 41
5.4.4 结果分析与临床建议 42
6 模型总结 **43**
6.1 模型的优点 43
6.2 模型的不足 43
6.3 未来的改进方向 43
参考文献 **44**
附录 A 问题一结果表格 **45**
附录 B 问题二结果表格 **51**
附录 C 问题三结果表格 **55**

1 问题背景与问题重述

1.1 问题背景

出血性脑卒中是一种危险性较高的脑血管事件，通常由于脑动脉瘤破裂、脑动脉异常等因素引起的脑实质内血管破裂而导致。它占据脑卒中发病率的 10-15%，具有高病死率和导致神经功能损伤的风险。

在出血性脑卒中的发展过程中，血肿扩张被认为是预后不良的主要危险因素之一。在短时间内，血肿可能会因脑组织受损、炎症反应等因素而逐渐扩大，导致颅内压力急剧升高，进而引发神经功能的进一步恶化，甚至危及患者生命。除此之外，血肿周围水肿的发生也引起了广泛的关注。血肿周围水肿可能导致脑组织受压，从而影响神经元的正常功能，导致进一步的损伤，增加患者的神经功能障碍。

鉴于出血性脑卒中患者的高危预后，早期识别和预测血肿扩张及血肿周围水肿的发生和发展对于个体化治疗和改善患者的生活质量具有至关重要的意义。然而，目前尚缺乏有效的方法来准确预测这些事件，从而采取及时的临床干预措施。随着医学影像技术的快速发展，我们有机会深入研究这些事件的发病机制和演化规律。医学影像可以提供宝贵的信息，包括血肿和水肿的体积、位置、形状以及灰度分布等。同时，人工智能技术的发展也为处理这些大规模医学影像数据提供了新的机会。因此，结合这些进展，研究如何根据临床和影像学信息，构建模型来预测患者的预后，以便更好地个性化治疗和改善出血性脑卒中患者的生活质量，具有重要的临床和科研价值。

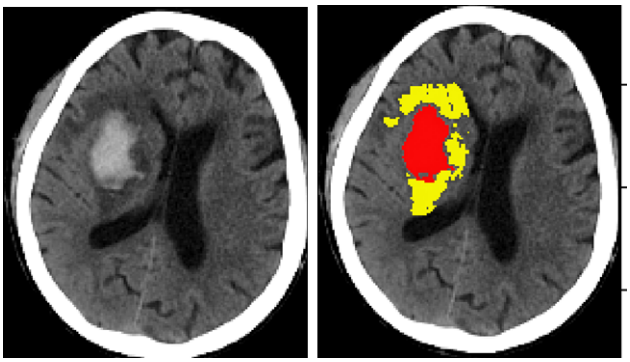


图 1.1　左图脑出血患者 CT 平扫，右图红色为血肿，黄色为血肿周围水肿

1.2 数据介绍

本题主要包含六个附件数据，各附件名称及内容见表1.1所示。具体来说，主要包含以下几个方面信息：

1、数据集

本题的数据集分为以下三部分：

表 1.1　数据文件介绍表

序号	文件名	内容
1	表 1-患者列表及临床信息	160 例（100 例训练数据集 +60 例独立测试数据集）出血性脑卒中患者的个人史、疾病史、发病及治疗相关信息、多次影像学检查（CT 平扫）结果及患者预后评估
2	表 2-患者影像信息血肿及水肿的体积及位置	患者影像学检查数据，包括每个时间点血肿和水肿的总体积及不同位置的占比
3	表 3-患者影像信息血肿及水肿的形状及灰度分布	患者影像学检查数据，包括每个时间点血肿和水肿的形状特征及灰度分布
4	表 4-答案文件	与建模目标相关的答案填写表格
5	附表 1-检索表格-流水号 vs 时间	患者的流水号与影像检查时间点的对应信息
6	附件 2-相关概念	出血性脑卒中临床相关概念，包括治疗方法、残差、患者预后、血肿及水肿影像特征等相关的概念和定义

（1）训练数据集（sub001 至 sub100）：包含 100 例患者的基本信息、首次及所有随访影像数据以及 90 天后的 mRS 评分。

（2）测试数据集 1（sub101 至 sub130）：包含 30 例患者的基本信息和首次影像数据，但不包含随访影像数据和 90 天 mRS 评分。

（3）测试数据集 2（sub131 至 sub160）：包含 30 例患者的基本信息、首次及所有随访影像数据，但不包含 90 天 mRS 评分。

2、目标变量

本题的目标变量包括以下两个：

（1）发病 48 小时内是否发生血肿扩张：这是一个二元分类变量，其中 1 表示发生血肿扩张，0 表示未发生。

（2）发病后 90 天 mRS 评分：mRS 评分是一个有序等级变量，范围从 0 到 6，用于评估患者在发病后 90 天的功能状态。其中，“0”表示“没有症状，没有残疾”；“1”表示“没有明显的残疾，能够独立进行日常活动”；“2”表示“有轻度残疾，能够自理，但在活动中存在一些限制”；“3”表示“有中度残疾，需要一定程度的帮助和照顾，但能够坐立或站立”；“4”表示“有中重度残疾，需要全天候照顾和帮助，无法行走或自理”，“5”表示“完全依赖他人，不能进行任何活动，床上活动有困难”，“6”表示“死亡”。

3、临床信息

患者的临床信息包括以下字段：

（1）ID：患者 ID。

（2）入院首次影像检查流水号：一个 14 位数字编码，前 8 位代表年月日，后 6 位为顺序编号，用于唯一标识影像检查。具体检查时间点可通过对应流水号在“附表 1-检索表格-流水号 vs 时间”中检索。

（3）年龄：患者的年龄（岁）。

（4）性别：患者的性别（男/女）。

（5）脑出血前 mRS 评分：患者脑出血前的 mRS 评分，范围从 0 到 6，是一个有序等级变量。

（6）高血压病史：是否有高血压病史，1 表示有，0 表示无。

（7）卒中病史：是否有卒中病史，1 表示有，0 表示无。

（8）糖尿病史：是否有糖尿病史，1 表示有，0 表示无。

（9）房颤史：是否有房颤史，1 表示有，0 表示无。

（10）冠心病史：是否有冠心病史，1 表示有，0 表示无。

（11）吸烟史：是否有吸烟史，1 表示有，0 表示无。

（12）饮酒史：是否有饮酒史，1 表示有，0 表示无。

4、发病相关特征

患者的发病相关特征包括以下两个字段：

（1）血压：包括收缩压和舒张压，单位为毫米汞柱（mmHg）。

（2）发病到首次影像检查时间间隔：单位为小时。

5、治疗相关特征

患者的治疗相关特征包括以下七个字段，用于记录是否接受特定的治疗措施：

（1）脑室引流：1 表示接受脑室引流治疗，0 表示未接受。

（2）止血治疗：1 表示接受止血治疗，0 表示未接受。

（3）降颅压治疗：1 表示接受降颅压治疗，0 表示未接受。

（4）降压治疗：1 表示接受降压治疗，0 表示未接受。

（5）镇静、镇痛治疗：1 表示接受镇静和镇痛治疗，0 表示未接受。

（6）止吐护胃：1 表示接受止吐护胃治疗，0 表示未接受。

（7）营养神经：1 表示接受营养神经治疗，0 表示未接受。

6、影像相关特征

患者的影像相关特征包括以下内容：

（1）血肿及水肿的体积和位置信息：这些信息包含在“表 2-患者影像信息血肿及水肿的体积及位置”中，包括每个时间点对应的血肿体积（HM_volume）、水肿体积（ED_volume）以及血肿和水肿在不同位置的占比，共 22 个特征。具体位置包括左右侧大

脑前动脉（ACA_L，ACA_R），左右侧大脑中动脉（MCA_L，MCA_R），左右侧大脑后动脉（PCA_L□PCA_R），左右侧脑桥/延髓（$Pons\_Medulla\_L$，$Pons\_Medulla\_R$），左右侧小脑（$Cerebellum\_L$，$Cerebellum\_R$）共十个不同位置，如图1.2。

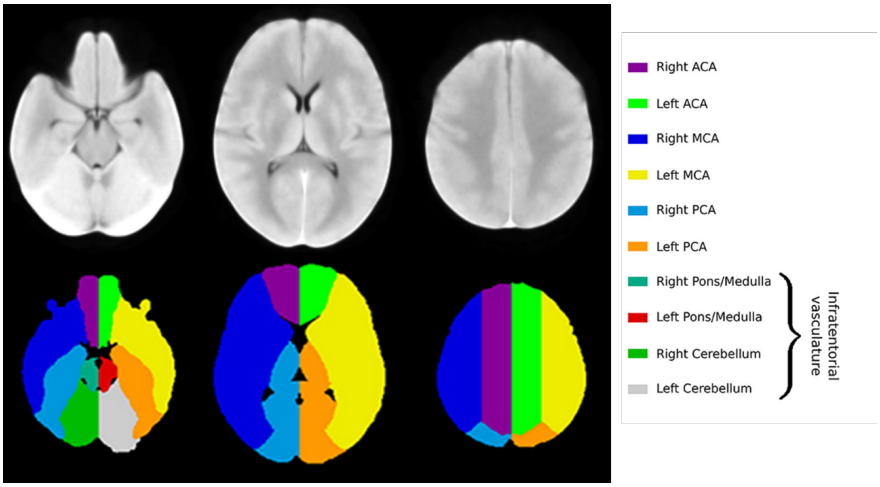


图 1.2　全脑位置

（2）血肿及水肿的形状及灰度分布信息：这些信息包含在“表 3-患者影像信息血肿及水肿的形状及灰度分布”中，该表分为两个不同标签页，“Hemo”对应血肿的形状及灰度分布信息，“ED”对应水肿的形状及灰度分布信息。每个时间点的血肿和水肿都包括 17 个灰度特征（基本的度量值，反映体素信号强度分布）以及 14 个形状特征（对目标区域三维形状描述），共计 31 个特征。

表 1.2　血肿及水肿的灰度和形状特征

灰度特征	10Percentile, 90Percentile, Energy, Entropy, InterquartileRange, Kurtosis, Maximum, MeanAbsoluteDeviation, Mean, Median, Minimum, Range, RobustMeanAbsoluteDeviation, RootMeanSquared, Skewness, Uniformity, Variance
形状特征	Elongation, Flatness, LeastAxisLength, MajorAxisLength, Maximum2DDiameterColumn, Maximum2DDiameterRow, Maximum2DDiameterSlice, Maximum3DDiameter, MeshVolume, MinorAxisLength, Sphericity, SurfaceArea, SurfaceVolumeRatio, VoxelVolume

1.3 问题重述

本题的建模目标是通过对表格提供的真实临床数据进行分析，研究出血性脑卒中患者的血肿扩张风险、血肿周围水肿发生及演进规律，并最终结合临床和影像信息，预测出血性脑卒中患者的临床预后。具体来说，需要解决以下三个问题：

问题一：血肿扩张概率预测建模及分析。

（1）基于前 100 个患者发病和多次影像检查时间，以及血肿体积前后变化，判断患者在发病后的 48 小时内是否发生血肿扩张事件，若发生则同时记录血肿扩张时间。

（2）基于患者的个人史、疾病史、发病相关信息，以及首次影像检查结果，构建模型预测所有患者发生血肿扩张的概率。

问题二：水肿体积进展建模及与治疗干预的关联关系探索。

（1）构建全体患者水肿体积随时间进展曲线，计算前 100 个患者真实值与拟合曲线之间的残差。

（2）探索患者水肿体积随时间进展模式的个体差异，构建不同亚组的水肿体积随时间进展曲线，同时计算前 100 个患者真实值与曲线之间的残差。

（3）分析不同治疗方法对水肿体积进展模式的影响。

（4）分析血肿体积、水肿体积以及治疗方法之间的关系。

问题三：出血性脑卒中患者预后预测建模及分析。

（1）利用前 100 个患者的个人史、疾病史、发病相关信息，以及首次影像结果，预测所有患者的 90 天 mRS 评分。

（2）利用前 100 个患者的已知临床、治疗信息以及影像结果，预测所有患者的 90 天 mRS 评分。

（3）分析出血性脑卒中患者的 90 天 mRS 评分与个人史、疾病史、治疗方法以及影像特征之间的关联关系，为临床决策提供建议。

2 模型假设与符号说明

2.1 模型假设

1. 患者数据独立同分布，即不同患者之间数据相互不干扰；
2. 患者特征之间存在一定关系，即不是完全独立；
3. 同一患者不同随访记录的影像特征和水肿体积与血肿体积存在因果联系，即不能将每次随访数据当作独立样本；
4. 水肿数据与血肿数据不完全独立；
5. 所给患者样本具有代表性，即通过本次训练学习的模型具有可推广性。

2.2 符号说明

表 2.3　符号说明

符号	意义
n	第 n 位患者
t_i	第 i 次影像检查的时间点
$\delta t_{i,j}$	从第 i 次到第 j 次影像检查的时间间隔
$VH_i(n)$	第 n 位患者第 i 次影像检查的血肿绝对体积
$VHr_i(n)$	第 n 位患者第 i 次影像检查的血肿相对体积
$p(n)$	第 n 位患者的血肿扩张概率
$VE_i(n)$	第 n 位患者第 i 次影像检查的水肿体积
$e(n)$	第 n 位患者水肿体积真实值与拟合曲线的残差
ρ	斯皮尔曼相关系数
P	斯皮尔曼相关系数的可信度
L	损失函数
Obj	目标函数

3 问题一：血肿扩张预测建模及分析

3.1 问题分析

问题一主要研究患者的血肿扩张，旨在探索出血性脑卒中患者发病后 48 小时内是否发生血肿扩张事件，并构建预测模型。该问题包括两个子问题：

（1）判断患者是否发生血肿扩张事件：

首先，我们关注了前 100 例患者（sub001 至 sub100），根据“表 1”中的入院首次影像检查流水号和发病到首次影像检查时间间隔，以及“表 2”中的各时间点流水号和对应的 HM_volume，来判断这些患者是否在发病后 48 小时内发生了血肿扩张事件。该判断基于血肿体积前后的变化，具体定义为后续检查相对于首次检查的绝对体积增加 $\geq 6mL$ 或相对体积增加 $\geq 33\%$。如果发生血肿扩张事件，则记录血肿扩张的具体时间。

（2）构建模型预测血肿扩张概率：

基于前 100 例患者（sub001 至 sub100）的个人史、疾病史、发病相关信息，以及“表 2”和“表 3”中的影像检查结果（仅包含对应患者首次影像检查记录），构建模型预测所有患者（sub001 至 sub160）发生血肿扩张的概率。这个模型将使用前 100 例患者的信息进行训练，然后将模型应用于所有患者，输出的概率值表示每位患者发生血肿扩张事件的可能性。

针对上述问题，提出如下求解思路：

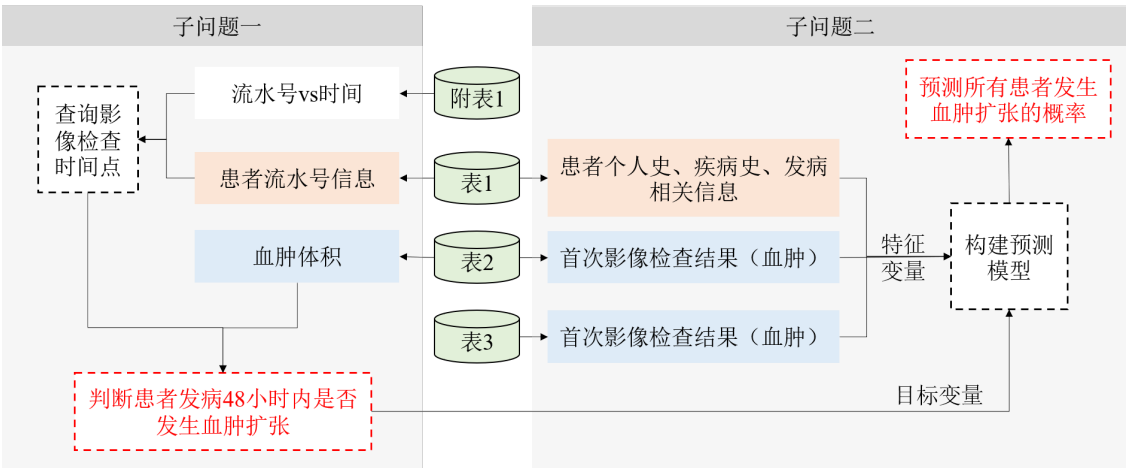


图 3.3　问题一求解思路

通过解决问题一的两个子问题，我们将更好地了解哪些因素与血肿扩张风险相关，并能够为临床决策提供预测患者是否会发生血肿扩张事件的信息。这对于及时干预和改善患者的治疗和管理具有重要的临床价值。

3.2 子问题一：48 小时血肿扩张判断

根据题干，首先提取“表 1”中入院首次影像检查流水号、发病到首次影像检查时间间隔的数据，以及“表 2”中各时间点流水号及对应的血肿体积 HM_volume 等特征数据。接着对患者 sub001 至 sub100 的信息进行遍历，并分别对第 n 位患者（$n = 1, 2, ..., 100$）执行以下步骤：

（1）通过流水号至“附表 1-检索表格-流水号 vs 时间”中查询相应影像检查时间点 $t_i(n)$，结合发病到首次影像时间间隔 $\delta t_0(n)$ 和后续影像检查时间，计算第 n 位患者发病到每次影像检查的时间间隔：

$$\delta t_{0,i}(n) = t_i(n) + \delta t_0(n), \quad i = 1, 2, 3 \tag{3.1}$$

经过数据处理和分析，我们发现所有患者第 4 次及以后的随访检查时间相对于首次检查时间均超过 48 小时。因此只需分析前 3 次影像检查结果。

（2）依次判断患者第 1 次到第 3 次的影像检查时间相对于发病的时间间隔是否小于 48 小时，即判断第 i 次检查时间是否满足：

$$\delta t_{0,i}(n) < 48, \quad i = 1, 2, 3 \tag{3.2}$$

若满足，则进入步骤 (3)。

（3）计算第 i 次检查相对于首次检查的血肿绝对体积的增加量 $\delta VH_i(n)$ 和相对体积增加量 $\delta VHr_i(n)$：

$$\begin{cases} \delta VH_i(n) = & VH_i(n) - VH_{i-1}(n), \quad i = 1, 2, 3 \\ \delta VHr_i(n) = & \dfrac{VHr_i(n) - VHr_{i-1}(n)}{VHr_{i-1}(n)}, \quad i = 1, 2, 3 \end{cases} \tag{3.3}$$

（4）判断第 n 位患者该次检查结果是否满足血肿扩张条件，即是否满足下述条件之一：

$$\begin{cases} \delta VH_i(n) \geq 6{\times}10^{-3}, & i = 1, 2, 3 \\ \delta VHr_i(n) \geq 33\%, & i = 1, 2, 3 \end{cases} \tag{3.4}$$

若满足，则说明该患者发病后 48 小时内发生了血肿扩张事件，在“表 4”的 C 字段中标记为 1，同时记录血肿扩张时间为 $\delta t_{0,i}(n)$，填入“表 4”D 字段。若没有发生血肿扩张事件，则在“表 4”的 C 字段中标记为 0。该记录的数据也可见本文的附录 A。

3.3 子问题二：血肿扩张概率预测

3.3.1 数据预处理

首先，我们从提供的数据集中选择和准备特征变量和目标变量。特征变量包括两部分，一部分是离散变量，即“表 1”中患者的个人史、疾病史、发病相关信息等，另一部分是连

续变量，即“表 2”和“表 3”中患者首次影像检查结果。目标变量是二元分类变量 HM，表示是否发生血肿扩张事件。

对于“表 1”中的患者信息，包括年龄、性别、脑出血前 mRS 评分、高血压病史、卒中病史、糖尿病史、房颤史、冠心病史、吸烟史、饮酒史、血压、脑室引流、止血治疗、降颅压治疗、降压治疗、镇静、镇痛治疗、止吐护胃、营养神经 18 个变量，分别记为 $A_1, A_2, ..., A_{18}$。其中，对年龄、性别和血压 3 个数据进行如下预处理：

（1）年龄 A_1：我们对患者的年龄数据进行了统计分析，计算了每个年龄值对应的患者数量，最终得到的统计图如图3.4所示。可以看到，年龄数据的最小值为 30 岁，最大值为 96 岁。基于此，我们将年龄分为间隔均匀的 10 段，并在后续的建模和分析中使用这些分段后的数据，以便更好地研究不同年龄段对血肿扩张的影响。

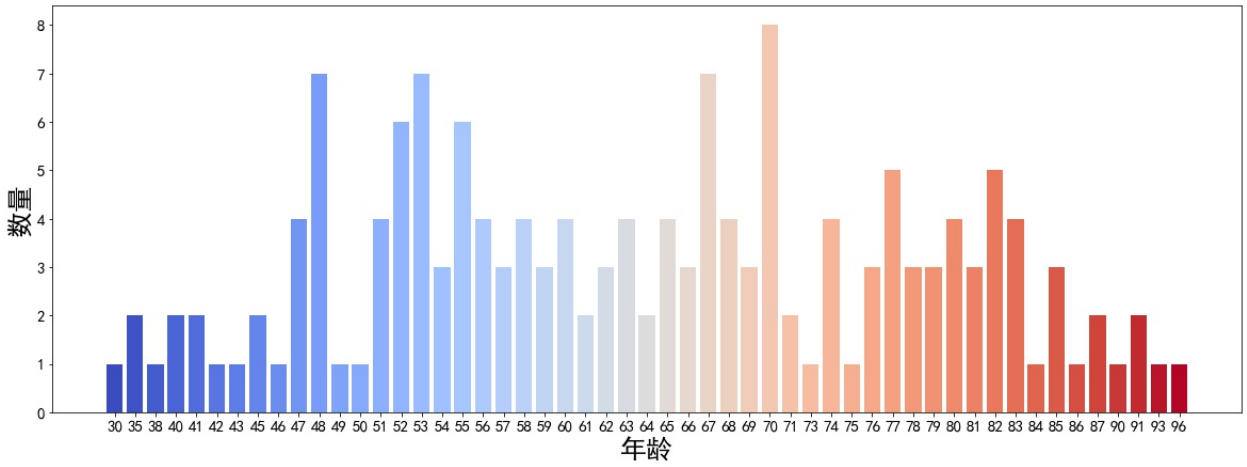


图 3.4　不同年龄的患者数量统计

（2）性别 A_2：若性别为男，则 $A_2 = 1$；若性别为女，则 $A_2 = 0$。

（3）血压 A_{11}：世界卫生组织对高血压的诊断标准有明确规定，正常人血压的收缩压在 $90mmHg$~$140mmHg$ 之间，舒张压在 $60mmHg$~$90mmHg$ 之间。根据该标准，将患者的血压分为三类。血压 $< 90/60mmHg$ 为低血压，记 $A_{11} = 0$；血压 $\geq 140/90mmHg$ 为高血压，记 $A_{11} = 1$；90~130/60~85$mmHg$ 属于正常血压，记 $A_{11} = 2$。

此外，在进行数据分析时，了解特征与目标变量之间的关系至关重要。在本题中，我们使用 Spearman 相关系数进行了“表 1”中特征变量与目标变量之间的相关性分析，可以帮助我们确定哪些特征对于预测目标变量最具影响力。

Spearman 相关系数是一种用于衡量两个变量之间的单调关系的非参数统计方法 [1]。与 Pearson 相关系数不同，Spearman 相关系数不依赖于数据的线性关系，而是基于变量的秩次进行计算，因此它对于非线性关系的探测更加敏感。通过计算相关系数，我们可以评估每个特征与目标变量之间的关系强度和方向。

最终计算得到的 Spearman 相关系数的热力图可视化如图3.5所示。其取值范围在-1 到

1 之间，其中 1 表示完全的正相关，-1 表示完全的负相关，0 表示没有相关性。同时在统计分析中，一般来说，相关系数绝对值大于 0.7 被认为是强相关，0.3 到 0.7 之间是中等相关，小于 0.3 则是弱相关。强相关的特征可能包含重复信息，可能不需要全部包含在模型中，以避免多重共线性问题。相反，中等或弱相关的特征可能在建模中具有独立的信息，对于预测目标变量是有益的。

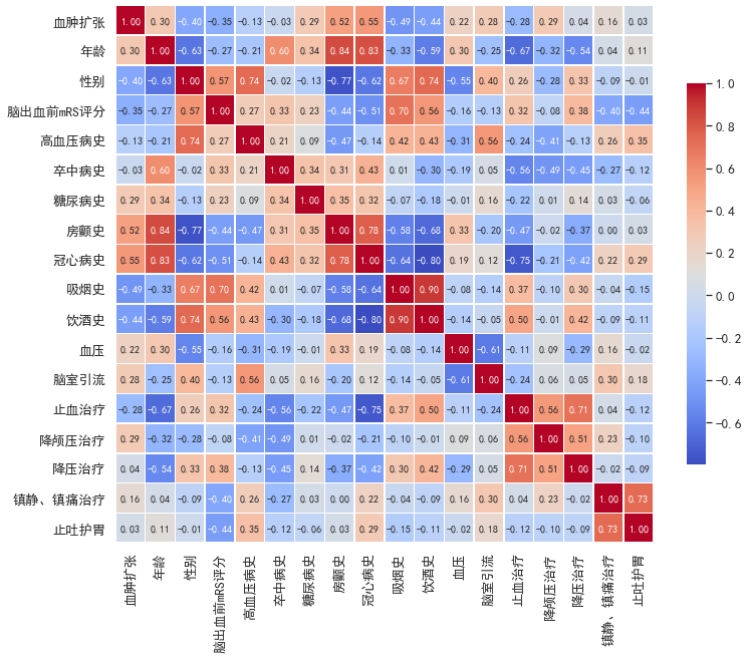


图 3.5　Spearman 相关系数热力图

通过观察，我们发现，本题的大多数特征之间的相关性被视为中等或弱相关，而只有少数特征之间存在较强的相关性。

3.3.2 主成分分析

由于该题中每位患者具有 71 维特征，特征数量过多会影响后续建模训练的效率，因此采用主成分分析（Principal Component Analysis，PCA）方法 [2] 对数据进行降维处理。

主成分分析法是一种降维技术，用于将高维数据映射到低维空间，同时保留数据的主要特征。PCA 的目标是找到数据中的主成分，这些主成分是原始特征的线性组合，可以最大程度地解释数据的方差。PCA 的核心思想是通过线性变换将数据映射到一个新的坐标系，使得数据在新坐标系下的方差最大化。这些新坐标轴被称为主成分，它们按照数据方差的降序排列，第一个主成分包含了最多的信息，第二个主成分包含了次多的信息，依此类推。通常，我们可以选择保留的主成分数量来控制降维的程度。

主成分分析的步骤主要包括：

（1）对数据进行标准化处理，这里将数据缩放到均值为 0、方差为 1 的范围内。数据标准化不仅能够消除不同特征之间的量纲差异，还可以使模型对于异常值或极端值的敏感

性降低，从而提高模型的稳定性，同时更容易让算法收敛，进而加速模型训练。

标准化方法如下公式所示：

$$x^* = \frac{x - \bar{x}}{\sigma} \tag{3.5}$$

式中，σ 为标准差，$\bar{x}$ 为 x 的平均值。

（2）计算标准化后的数据的协方差矩阵：

$$R = (r_{ij})_{p\times p} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \tag{3.6}$$

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{ij} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{ij} - \bar{x}_j)^2}} \tag{3.7}$$

协方差矩阵描述了数据特征之间的线性关系。

（3）对协方差矩阵进行特征值分解，计算特征值和对应的特征向量：

$$\begin{cases} \lambda = (\lambda_1, \lambda_2, \ldots \lambda_p) \\ a_i = (a_{i1}, a_{i2}, \ldots a_{ip}), \quad i = 1, 2, \ldots p \end{cases} \tag{3.8}$$

特征值表示了数据中的方差，特征向量是协方差矩阵的特征轴，每个特征向量对应一个特征值。特征值和特征向量是 PCA 的核心输出。

（4）根据问题的要求或数据的性质，选择要保留的主成分数量。通常可以根据特征值的大小来决定。较大的特征值表示了较多的方差，因此可以选择保留那些特征值较大的主成分。

值得注意的是，在本题中，由于不同患者影像检查次数不同，因此存在大量缺失值数据。常见的缺失值处理方法有均值填充、中位数填充、最大值填充、“0”填充等。经过测试对比，我们发现填充“0”的准确率最高，推测是因为在该题中“0”代表没有该影像特征。

接着，我们进行 PCA 处理，选择保留原数据 99% 的特征，将特征数据降到了 40 维。

3.3.3 模型构建

为了充分利用数据并评估模型的性能，我们采用了三种不同的建模方法：随机森林、Logistic 回归和 XGBoost。

1、Logistic 回归

Logistic 回归分析属于非线性回归，它是研究因变量为二项分类或多项分类结果与某些影响因素之间关系的一种多重回归分析方法 [3]。该模型的核心思想在于通过建立一个

逻辑函数（Logistic 函数）来估计一个样本属于某一类别的概率，即将输入的特征映射到一个概率输出，用于描述事件发生的可能性。

Logistic 回归模型的逻辑函数采用了 S 形曲线，其输出值在 0 到 1 之间，可以表示为概率。该函数具有良好的性质，使得 Logistic 回归能够灵活地适应各种复杂的分类问题。模型的训练过程主要涉及到参数的估计，通常采用了最大似然估计方法来确定模型的参数。

2、随机森林

随机森林 [4] 是一种强大而广泛应用的机器学习算法，它在分类和回归问题中都表现出色。

随机森林模型是由多棵决策树组成的。决策树 [5] 是一种树状结构，用于将数据集分成不同的类别或进行回归分析。它由节点、分支、叶子节点组成。决策树的构建过程是根据数据特征来选择最佳的划分点，以将数据分成最纯净的子集。这个过程不断重复，直到达到某个停止条件（如树的深度达到一定值或节点中的样本数量小于阈值）。随机森林的核心思想是集成学习，它通过组合多个弱学习器来构建一个强大的学习器。在随机森林中，这些弱学习器就是决策树。通过构建多个不同的决策树，随机森林可以减小过拟合的风险，提高模型的鲁棒性和准确性。

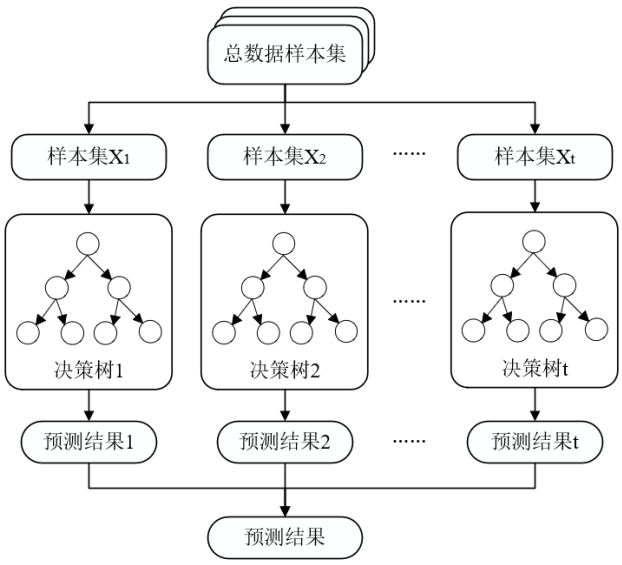


图 3.6　随机森林算法流程图

构建随机森林的过程包括以下步骤：

（1）样本的随机抽样：首先，从总体样本容量为 M 的数据集中，采用有放回的方式随机抽取 N 次，每次抽取 1 个，最终形成了 N 个新的、具有随机性的样本集。选择好了的 N 个样本用来训练一个决策树，作为决策树根节点处的样本。这个随机抽样过程确保了每个决策树的训练样本都具有差异性，从而增加了随机森林的多样性和鲁棒性。

（2）属性的随机选择：在构建每个决策树的过程中，每个节点需要选择一个属性来进

行分裂。假设总共有 M 个属性可供选择，在每个节点的分裂过程中，随机选择 m 个属性（通常情况下 $m << M$），然后从这 m 个属性中采用某种策略（比如说信息增益）来选择 1 个属性作为该节点的分裂属性。这个随机属性选择的过程有助于决策树的多样性，避免了单一属性的主导性。

（3）决策树的构建：在每个决策树的根节点处，使用步骤 (1) 中选定的样本集。然后，按照步骤 (2) 中随机选择的属性进行分裂。这个过程会一直进行，直到达到某个停止条件。需要注意的是，在决策树构建过程中，并没有进行剪枝操作，允许每个树生长到足够深。

（4）集成多个决策树：通过重复上述步骤，可以得到多个决策树。在分类问题中，通过多数投票的方式确定最终的分类结果。在回归问题中，将多个决策树的预测结果取平均作为最终的回归结果。这个集成的过程有助于提高模型的准确性和泛化能力。

3、XGBoost

XGBoost (Extreme Gradient Boosting) 算法 [6] 是一个可扩展的分布式梯度提升决策树 (GBDT) 机器学习库。XGBoost 支持并行树提升，是用于回归或分类问题的领先机器学习模型。

XGBoost 的核心思想在于将弱分类器（通常是决策树）组合成一个强大的集成模型。其独特之处在于引入了梯度提升算法，并采用了一系列创新性的技术来提高模型的准确性和效率。该算法的工作方式如下：首先，XGBoost 以一个弱分类器（单棵决策树）作为初始模型，然后计算每个样本点的残差（实际值与当前模型预测值之间的差异）。接着，它训练一个新的决策树来拟合这些残差，以纠正先前模型的错误。这个过程不断迭代，每一轮都关注之前模型的错误，以使得每个新模型都更加精确。

对于包含 n 条 m 维的数据集，XGBoost 模型可表示为：

$$\hat{y}_i = \sum_{k=1}^{K} f_k\left(x_i\right), \quad f_k \in F \quad (i = 1, 2, \ldots n) \tag{3.9}$$

其中，$F = \left\{f(x) = w_{q(x)}\right\}\left(q : R^m \to \{1, 2, \ldots T\}, w \in R^T\right)$ 是 CART 决策树结构集合，q 为样本映射到叶子节点的树结构，T 为叶子节点数，w 为叶节点的实数分数。

一般而言，损失函数描述的是预测值 y 与真实值 $\hat{y}$ 之间的关系。这里使用平方损失函数，对于 n 个样本，可以写成：

$$L = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) \tag{3.10}$$

更进一步，目标函数可以写成：

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \Omega\left(f_t\right) \tag{3.11}$$

其中 Ω 代表基模型的复杂度，若基模型是树模型，则树的深度、叶子节点数等指标均可以反映树的复杂度。

3.3.4 模型训练

1、数据集划分

对于 Logistic 回归，采用 K 折交叉验证 (k-fold cross-validation) 的方法将数据集划分成 K 等分，依次取每一份作为测试集，剩下的 $K-1$ 份为训练集 [7]。交叉验证重复 K 次，取 K 次准确率的平均值作为最终模型的评价指标。它能够保证每个子样本都参与训练且都被测试，降低泛化误差，有效避免过拟合和欠拟合状态的发生。其中 K 值的选择根据实际情况调节。对于随机森林和 XGBoost 算法，随机选取 80% 的数据作为训练集，20% 的数据作为测试集。

2、超参数选择

超参数是机器学习算法中的配置选项，它们不是由模型自动学习的，而是由数据科学家或研究人员在模型训练之前手动设置的。超参数的选择可以显著影响模型的性能，因此在训练之前，我们需要对每个模型进行最优化参数选择，这里使用 GridSearch 方法。

GridSearch（网格搜索）是一种用于系统地搜索机器学习模型的超参数组合以找到最佳组合的方法。其基本思想是在预定义的一组超参数值中进行穷举搜索，对每个超参数组合都训练模型并评估性能。

在模型调参之前，我们需要明确定义每个模型的超参数搜索空间。对于 XGBoost 和随机森林，我们关注了以下超参数：学习率（$learning\_rate$）、树的深度（max_depth）、树的数量（$n\_estimators$）、最小叶节点样本数（$min\_samples\_leaf$）、最大特征数（$max\_features$）。对于 Logistic 回归，我们主要关注正则化参数 C 的值。

使用 GridSearch 方法后，我们找到了每个模型的最佳超参数组合，代入模型进行训练。

3.3.5 模型对比

模型训练结束后，我们分别计算了三个模型对测试集数据预测的准确率，结果如表3.4所示。

表 3.4　三种模型准确率对比

模型	测试集上的准确率
Logistic	76%
随机森林	85%
XGBoost	90%

由表可知，XGBoost 的准确率最高，说明该模型预测效果最好。根据题目要求，使用 XGBoost 算法对所有患者（sub001 至 sub160）发生血肿扩张的概率进行预测，最终得到的结果填写在“表 4”的 E 字段（血肿扩张预测概率）中。同时，也可见本文的附录 A。

4 问题二：水肿体积进展建模及分析

4.1 问题分析

问题二旨在研究出血性脑卒中患者的血肿周围水肿的发生及进展模式，以及探索治疗干预与水肿进展之间的关联关系。该问题包括以下四个子问题：

（1）构建全体患者的水肿体积随时间进展曲线，并计算残差：

为了了解整体趋势，首先，我们针对前 100 位患者的水肿体积 ED_volume 与重复检查时间点构建了一条水肿体积随时间进展的曲线。在这个过程中，我们考察了水肿体积的动态变化，x 轴表示从发病到影像检查的时间，y 轴表示水肿体积。然后，我们计算了这一曲线上每个时间点的真实水肿体积与拟合曲线之间的残差。残差反映了模型的拟合效果，正值表示拟合值高估了水肿体积，负值表示低估了水肿体积。

（2）探索不同人群的水肿体积随时间进展模式及残差：

在这一部分需要研究不同患者群体之间水肿体积随时间进展模式的个体差异。我们通过对患者水肿体积变化的线条趋势进行聚类分析，归纳特点，并据此将患者分为四个亚组，每个亚组代表一组患者，具有相似的水肿体积随时间进展特点。对于每个亚组，我们构建了相应的水肿体积随时间进展曲线，并计算了各个亚组拟合曲线的残差以及整体残差，与第一问展开了对比分析，验证了分类模型的准确性。

（3）分析不同治疗方法对水肿体积进展的影响：

这一部分旨在探讨不同治疗方法对水肿体积进展模式的影响。我们将分析患者接受不同治疗方法的情况，是否与水肿体积的进展模式存在相关性。分析方法包括使用决策树算法得到各个治疗组对进展模式的特征重要性，以及使用统计学方法来可视化并进一步分析相关性。

（4）分析血肿体积、水肿体积及治疗方法之间的关系：

我们将通过整体相关性分析以及图表分析探讨分析血肿体积、水肿体积以及治疗方法这些因素之间是否存在相互影响或关联。这一分析可以帮助我们更好地理解治疗方法如何影响水肿的发生和进展，为临床决策提供有价值的信息。

4.2 子问题一：全体患者水肿体积进展曲线

4.2.1 水肿体积随时间变化分布

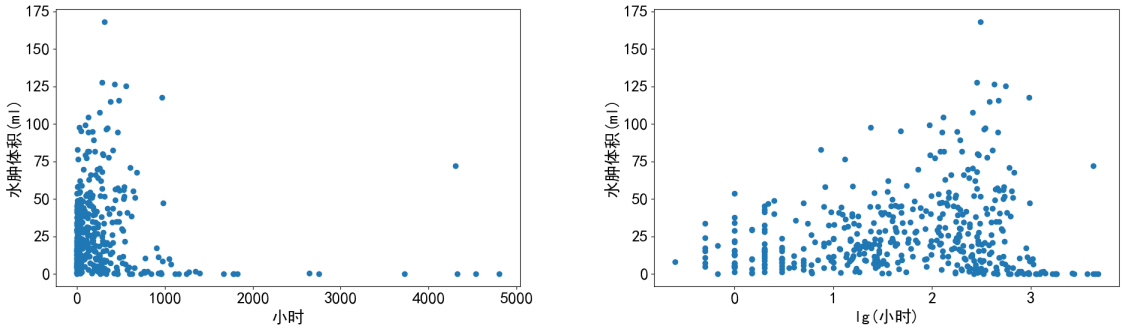
为了更好地理解水肿体积发生和进展趋势，我们采用数据可视化，首先制作了水肿体积随时间变化的分布散点图。

初始的散点图如图4.7(a) 所示，展示了横坐标（时间）和纵坐标（水肿体积）之间的关系。然而，通过观察，我们注意到横坐标上的数据点相对较为集中，这可能导致在线性比例下的散点图不够清晰，难以准确捕捉水肿体积随时间的趋势。这种情况下，我们需要

采取一些数据处理步骤，以更好地呈现数据的模式。

为了解决这个问题，我们决定对横坐标（时间）进行对数转换。对数转换是一种常见的数据变换方法，特别适用于数据范围广泛或呈指数增长的情况。在这里，取对数的目的是拉伸横坐标的刻度，使得时间的变化更平滑地映射到图表上。

经过对数转换后，我们重新绘制了水肿体积随时间变化的散点图，如图4.7(b) 所示。与初始的散点图相比，新的图表更容易拟合，更清晰地展示了水肿体积的趋势。这种数据处理方法有助于凸显时间对水肿体积的影响，使我们能够更准确地观察到水肿的发生和进展情况。为更好的展示图的结果，后面将直接对横坐标（小时）取对数处理，并不再说明。



(a) 横坐标不取对数　　(b) 横坐标取对数

图 4.7　水肿体积随时间变化分布散点图

4.2.2 水肿体积随时间进展曲线拟合

1、RBF 核函数拟合

径向基函数（Radial Basis Function，RBF）核函数是支持向量机（Support Vector Machine，SVM）中常用的核函数之一，也被广泛应用于机器学习和数据挖掘领域。RBF 核函数的特点在于其出色的非线性映射能力，能够有效地处理线性不可分的数据，并将其映射到高维特征空间，从而实现了更好的分类性能。

RBF 核函数的核心思想源自径向基函数，这是一个以原点为中心，从中心点向外扩散的函数，具有类似钟形的形状。其数学表达式如下：

$$K\left(x, x'\right) = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{4.12}$$

其中，x 和 x' 分别表示输入样本的两个特征向量，$\|x - x'\|$ 是这两个向量之间的欧氏距离，σ 是一个控制函数形状的参数。σ 的选择会直接影响到分类器的性能——较小的 σ 值会导致核函数的扩散范围较小，模型会更加关注训练样本的局部特征，容易过拟合；而较大的 σ 值则会导致核函数的扩散范围较大，模型更关注整体特征，容易欠拟合。因此，选择合适的 σ 参数是应用 RBF 核函数时需要仔细调优的一部分。

总之，RBF 核函数的关键在于它通过将输入数据映射到高维特征空间，实现线性不可分问题的线性可分化，为机器学习领域提供了重要的工具和方法。

在本题中，使用 RBF 核函数拟合得到的前 100 个患者（sub001 至 sub100）水肿体积随时间变化拟合曲线如图4.8所示。

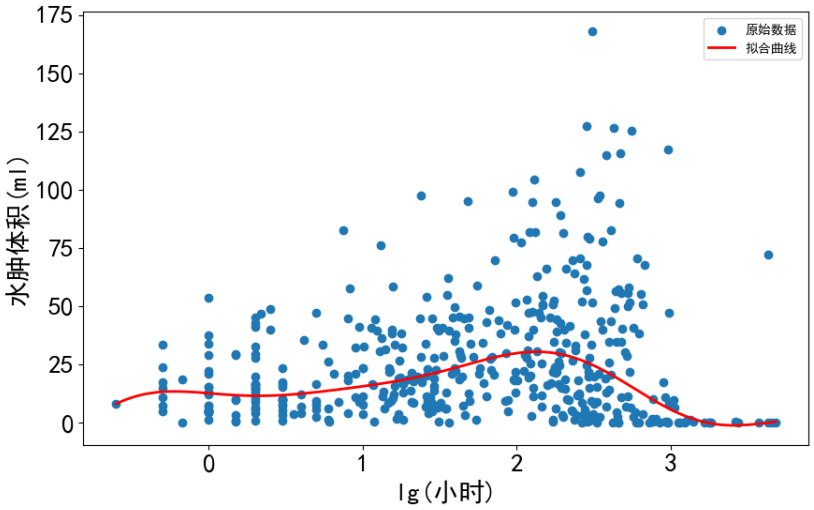


图 4.8　水肿体积随时间变化拟合曲线（核函数拟合）

2、样条插值拟合

样条插值是一种常用的数值分析方法，用于拟合或逼近一组离散数据点，以生成一个光滑的曲线或曲面。它在数据重建、函数逼近和曲线拟合等领域具有广泛的应用。样条插值的核心思想是将数据区间分割成若干小段，每一段用一个低阶多项式来逼近，然后将这些多项式组合成一个整体的光滑函数。

样条插值方法的优势之一是它们不仅适用于均匀分布的数据点，还适用于不均匀分布的数据点，因为它们在每一段内都使用多项式逼近，从而能够更好地处理数据点密集和稀疏的情况。

使用 25 次样条插值拟合得到的前 100 个患者（sub001 至 sub100）的水肿体积随时间变化曲线如图4.9所示。

3、残差计算

考虑到同一名患者有可能在这多次检测结果中与拟合曲线出现正负误差，导致最终的残差和变小，因此对每一个残差进行绝对值处理，最终根据检测次数求平均得到每一位患者与拟合曲线的残差值。因此，我们基于以下公式计算单个残差：

$$e(n)=\frac{\sum_{i=1}^{m(n)}|y_i(n)-\hat{y}_i(n)|}{m(n)} \tag{4.13}$$

其中 $m(n)$ 表示第 n 位患者的影像检查次数。

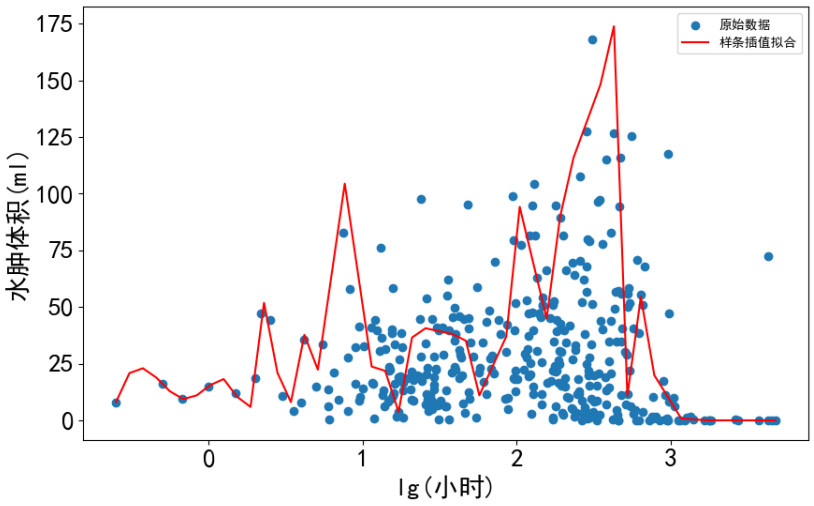


图 4.9　水肿体积随时间变化拟合曲线（样条插值拟合）

平均残差是所有残差的平均值，可以表示模型整体的预测误差。计算平均残差有助于评估模型的拟合程度，如果平均残差接近零，表示你的模型对数据的拟合较好；如果平均残差远离零，说明模型可能存在较大的预测误差。本题中使用以下公式计算，其中 100 表示患者数量：

$$\bar{e} = \frac{\sum_{n=1}^{100} e(n)}{100} \tag{4.14}$$

4、拟合曲线对比

根据残差的定义，可以计算得到核函数和 25 次样条插值的平均残差结果如下：

表 4.5　平均残差结果

拟合方法	平均残差值 (ml)
核函数拟合	20.15
样条插值拟合	-44.20

结果显示，RBF 核函数拟合在描述水肿体积随时间的变化趋势方面表现更出色，这意味着它更准确地捕捉了数据的特征。这一发现强调了在研究水肿进展模式时选择合适的数学拟合方法的重要性。通过采用 RBF 核函数，我们能够更好地理解水肿体积的动态变化，这有助于我们更全面地了解出血性脑卒中患者的病情发展。这些发现将在后续的问题分析和临床决策中提供宝贵的信息和支持。

因此，本小题选择核函数拟合曲线进行后续的残差计算和分析。

计算得到的单个残差将被记录在“表 4”的 F 字段（残差（全体））中，同时对这些数据可视化，得到的折线图如图4.10所示。

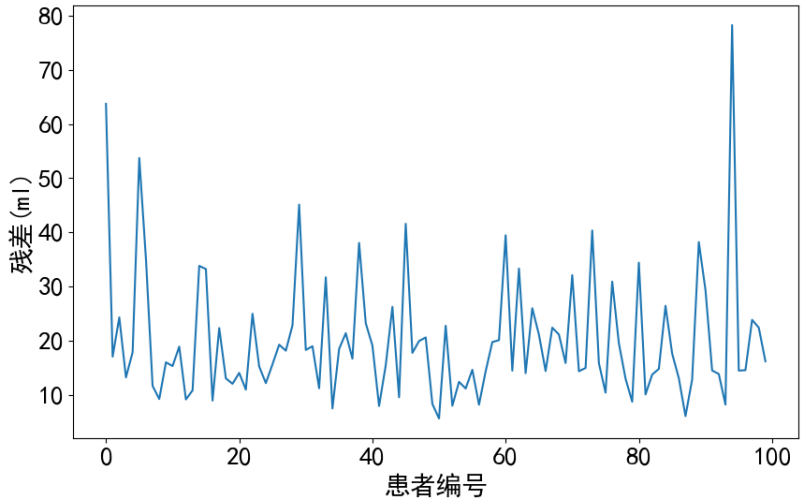


图 4.10　核函数拟合全体患者残差图

4.3 子问题二：不同人群的水肿体积进展曲线

在这一部分，我们将探讨患者水肿体积随时间进展的个体差异，以更深入地理解血肿周围水肿的发生和进展模式。

首先，我们依旧选用图4.7(b) 的水肿体积随时间的分布散点图进行分析，即对横坐标（时间）取对数，使数据更符合模型假设。

接着，我们使用不同的数据分组和曲线拟合方法来处理这些数据。考虑到共有一百位患者的水肿体积随时间进展的方式，我们将所有患者的水肿体积变化组合成折线，并进行线条走向的聚类分析。所有患者的水肿变化趋势如下：

我们将所有线条具化为三段大小：初始斜率大小、中间斜率大小、末尾斜率大小，形成趋势的三维点坐标，通过 Mean Shift 算法对三维线条趋势点进行聚类，并估计密度，生成相应聚类中心。

通过对聚类中心的分析，我们最终将患者分成两大部分：一部分呈现明显的水肿体积变化（总体趋势呈现变化），另一部分水肿体积相对稳定（总体趋势不变）。对于呈现变化的部分，我们进一步将其分为三个亚组：水肿体积总趋势上升组、水肿体积总趋势下降组、水肿体积先上升后下降组。

最终，我们将患者分为四个不同的随时间进展的水肿体积变化趋势组，包括：

（1）亚组 1——水肿体积稳定在较低水平组：这些患者的水肿体积相对稳定，没有显著的变化。

（2）亚组 2——水肿体积总趋势下降组：这组患者水肿体积呈逐渐减小的趋势。

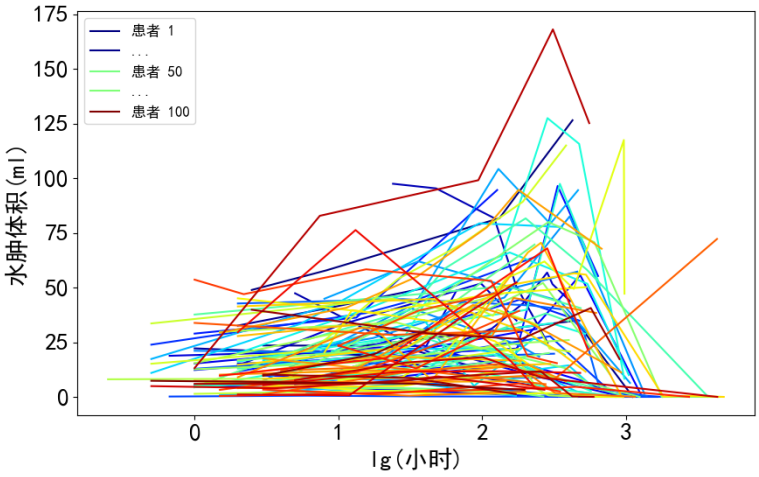


图 4.11　全体患者的水肿变化趋势线条

（3）亚组 3——水肿体积先增加后减少组：在一段时间内，这些患者的水肿体积增加，然后开始逐渐减小。

（4）亚组 4——水肿体积总趋势上升组：这组患者的水肿体积呈逐渐增加的趋势。

针对每个亚组，我们分别进行了水肿体积随时间的曲线拟合，并在一张综合的散点图中呈现了四条不同的拟合曲线，如图4.12所示。其中，蓝色为水肿体积稳定在较低水平组，红色为水肿体积总趋势下降组，绿色为水肿体积先增加后减少组，黄色为水肿体积总趋势上升组。

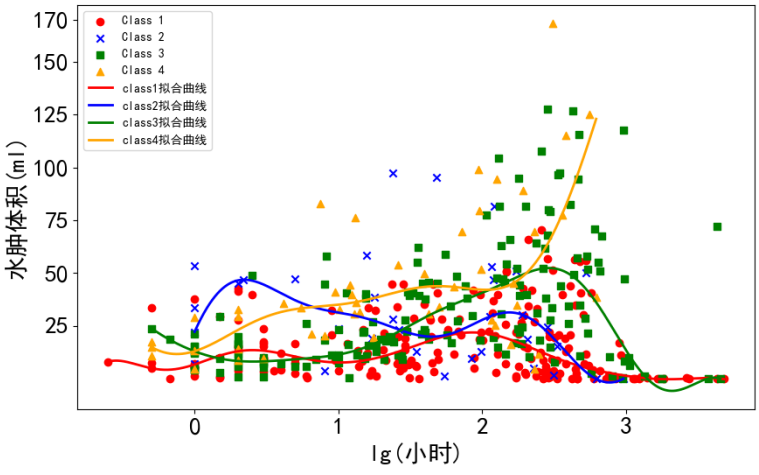
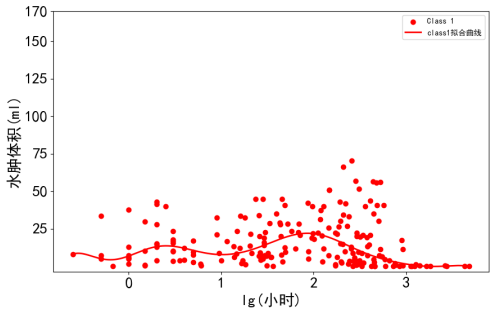
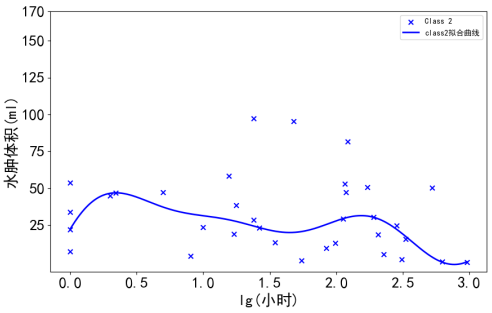


图 4.12　四分类下的水肿随时间变化体积分布和拟合曲线

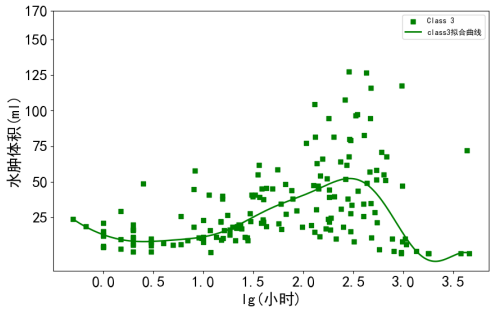
此外，我们还绘制了四个子图，每个子图仅包含一个亚组的数据分布和拟合曲线，如图4.13。我们认为这种方式有助于将不同趋势的数据分布和拟合结果更清晰地展示出来，减少混淆的同时更易于理解和解释。同时也有助于突出不同水肿进展模式之间的区别，凸



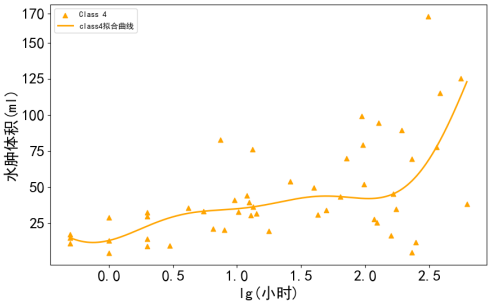
(a) 水肿体积稳定在较低水平组



(b) 水肿体积总趋势下降组

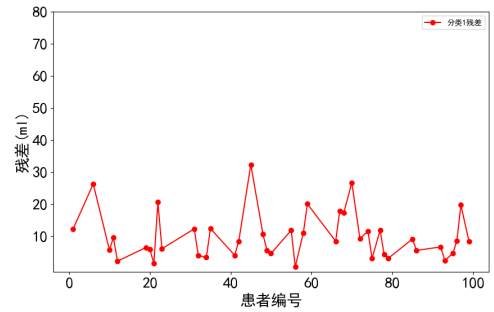


(c) 水肿体积先增加后减少组

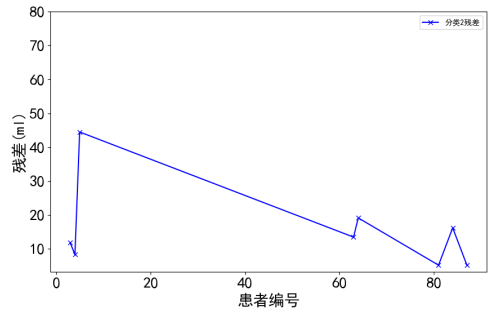


(d) 水肿体积总趋势上升组

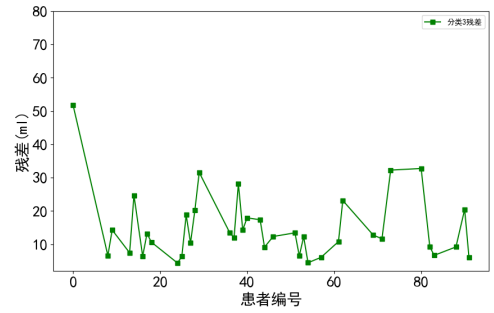
图 4.13　不同亚组的水肿体积随时间变化分布和拟合曲线



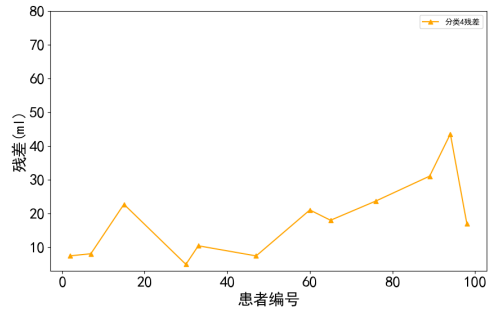
(a) 水肿体积稳定在较低水平组



(b) 水肿体积总趋势下降组



(c) 水肿体积先增加后减少组



(d) 水肿体积总趋势上升组

图 4.14　四条拟合曲线的独立残差折线图

显了患者之间的个体差异，从而更全面地揭示了水肿体积随时间的变化规律。

为了更精确地评估模型的拟合效果，对于每条拟合曲线，我们计算了四条曲线的独立残差并绘制成折线图，如图4.14所示，可以看到大部分患者的残差较小，说明四条曲线拟合结果较好。这些残差将被记录在“表 4”的 G 字段（残差（亚组））中，并在“表 4”的 H 字段（所属亚组）中指明患者所属的亚组。同时，也可见本文的附录 B。

对于每个亚组，我们计算了曲线拟合的平均残差值和标准差值，结果如表4.6所示。

表 4.6　不同亚组的平均残差及标准差

	平均残差值 (ml)	标准差 (ml)
亚组 1	9.9	7.2
亚组 2	15.5	11.9
亚组 3	15.0	9.8
亚组 4	17.9	11.0

从上表可以分析得出，亚组 1（总趋势下降组）拟合效果是最佳的，因为该分类依据是处于稳定状态。其他的亚组残差值以及标准差会比亚组 1 较大，这是因为这三个亚组波动比较明显，会有较多远离拟合曲线的点。但总体来说，对患者水肿体积随时间的发展趋势分成四组后，整体的拟合程度提升较好。

此外，我们绘制了分类后的所有患者的残差折线图，如图4.15所示。

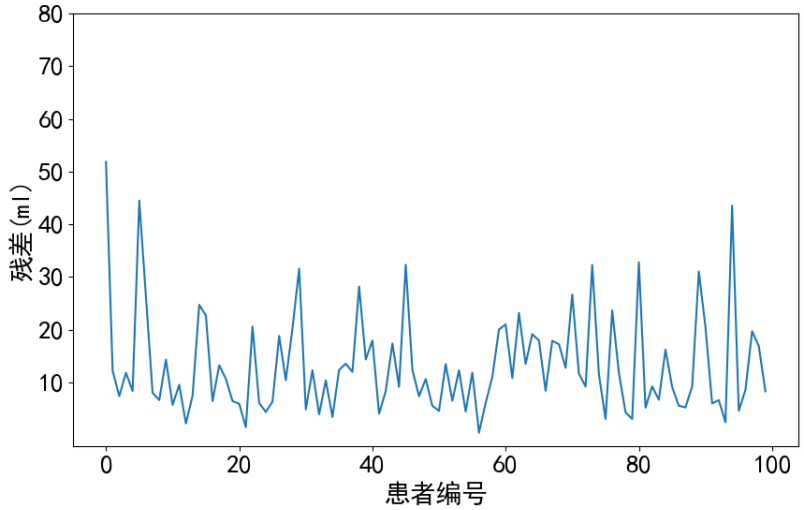


图 4.15　分类后的全体患者残差图

该图与子问题一中的残差图对比如图4.16所示。

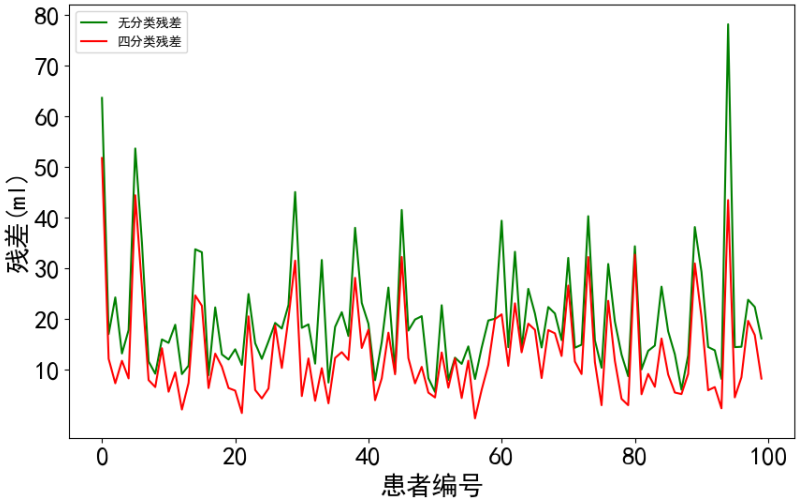


图 4.16　无分类残差与四分类残差对比

由于核函数对非线性数据的拟合效果较好，我们对四种种类的患者数据都采用了与上一问相同的 RBF 核函数拟合方法，因此在残差图上与无分类的波动趋势较为相似。对于分类后的拟合结果，我们可以计算出平均残差和提升效果如表4.7。

表 4.7　平均残差结果对比

	平均残差 (ml)	提升比例
分类拟合	13.24	34.3%
整体拟合	20.15	

可见，对患者水肿体积随时间变化分类后，平均残差有较好的提升，拟合曲线更接近患者的真实情况，获得了良好的拟合效果。

上述分析结果为我们提供了关于不同患者水肿体积变化模式的更深入理解，为进一步探索与治疗干预和水肿进展的关联关系奠定了基础。这也为个性化治疗策略的制定提供了有益的信息。

4.4 子问题三：不同治疗方法对水肿体积进展的影响

4.4.1 决策树分析

为了探究患者的七种治疗方式对水肿体积随时间变化的发展模式，首先考虑使用决策树来分析七个变量对一个结果的影响。

由于子问题二中使用 1，2，3，4 的方法代表分类结果，数字之间存在着天然的距离，为了保证较好的分析结果，考虑使用独热化方式对分类结果事先进行编码。

独热编码（One-Hot Encoding）是一种常用于处理分类数据的编码技术，它将一个具有多个类别的离散特征转换为二进制向量形式，以便机器学习模型能够更好地理解和处理这些特征。这个编码方法的核心思想是将每个类别都映射到一个独立的二进制特征，其中只有一个特征位为 1，其余特征位都为 0，以表示该类别的存在或不存在。它能够保留类别之间的无序性，并将其转化为算法可接受的形式。

具体来说，对于每个分类特征，首先需要确定该特征有多少个不同的类别（取值）。然后，针对每个类别，创建一个二进制特征，将其命名为类别名称，并在对应类别的位置上设置为 1，其他位置设置为 0。这样就形成了一个稀疏的二进制向量。

独热编码的主要优点之一是它不引入类别之间的顺序关系，因此适用于名义型（nominal）或无序分类数据。它还可以防止模型错误地学习到不正确的关联性。此外，独热编码的结果在一定程度上具有可解释性，模型可以更容易地理解和解释特征的影响。

最后编码结果如表4.8所示。

表 4.8　独热编码结果

	独热化二进制编码
亚组 1	1000
亚组 2	0100
亚组 3	0010
亚组 4	0001

为了选择最佳的决策树参数，使用 Grid Search 方法来寻找最佳超参数组合。定义决策树的超参数为最大深度、内部节点分裂所需的最小样本数、叶子节点所需的最下样本数，搜索后得到的最佳参数分别为：不限制深度 (None)、1、2。

使用上述参数进行决策树分析，对数据集采取三七分的策略，即 70% 为训练集，30% 为测试集。最终在测试集中根据患者的治疗方案预测患者水肿体积变化模式，预测准确率为 46.7%。

其中一次分析得到的混淆矩阵如下：

$$CM = \begin{bmatrix} 4 & 0 & 6 & 1 \\ 0 & 1 & 0 & 0 \\ 6 & 0 & 8 & 0 \\ 0 & 0 & 3 & 1 \end{bmatrix} \tag{4.15}$$

根据混淆矩阵的定义，如表4.9所示，可以得知预测的错误主要发生在实际类别 1 预测类别 3、实际类别 3 预测类别 1 上。这可能是由于不同的治疗方式对这两类水肿体积发展的模式影响相似。

表 4.9　混淆矩阵定义

	预测类别 1	预测类别 2	预测类别 3	预测类别 4
实际类别 1	True Positives(TP1)	False Negatives(FN1)	False Negatives(FN1)	False Negatives(FN1)
实际类别 2	False Positives(FP2)	True Positives(TP2)	False Negatives(FN2)	False Negatives(FN2)
实际类别 3	False Positives(FP3)	False Positives(FP3)	True Positives(TP3)	False Negatives(FN3)
实际类别 4	False Positives(FP4)	False Positives(FP4)	False Positives(FP4)	True Positives(TP4)

进一步定性七种治疗方式对区分出水肿体积发展模式的影响性大小，可从决策树中输出七种治疗方式的特征重要性如表4.10所示：

表 4.10　七种治疗方式的特征重要性

脑室引流	止血治疗	降颅压治疗	降压治疗	镇静、镇痛治疗	止吐护胃	营养神经
0.211	0.123	0.119	0.101	0.120	0.092	0.056

根据特征重要性排序可得影响力从大到小分别为：脑室引流，止血治疗，镇静、镇痛治疗，止血治疗，降颅压治疗，降压治疗，营养神经。其中最大的为脑室引流，但也仅仅处于 0.211，属于弱相关的水平。其中营养神经对水肿体积发展模式影响最小。

4.4.2 统计学分析

首先对所有患者而言，存在着四种发展模式之间的比例关系。为了更直观的分析不同的治疗方式对患者水肿体积发展模式的影响，得到每个治疗方式下患者发展模式的比例分布如图4.17所示。

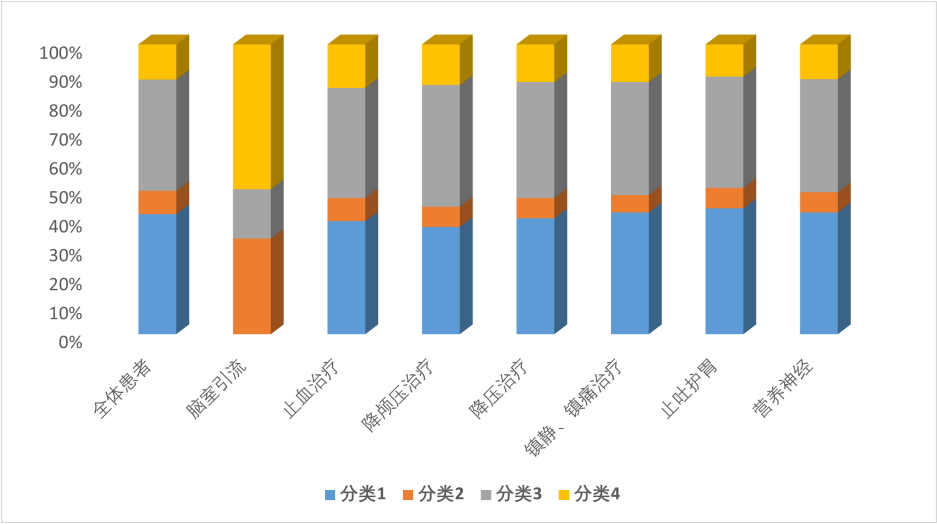


图 4.17　治疗方法所占分类比例图

由图可分析得到，脑室引流是比例分布差异最大的治疗方式，其不存在亚组 1（稳定低水平）的患者，同时其他三种分类方式的分布也与其他治疗方式有较大差异。其余六种治疗方案的所占发展模式比例都类似，这有很可能是由于其余六种方式的采取比例都较高，均有 3/4 以上的患者采取。

4.4.3 总结

虽然在图4.17中脑室引流的发展模式与其他治疗方式比例分布差异非常大，但 100 个患者中仅有六人采用了该治疗方式，因此计算其相关系数时仍处于弱相关的水平。同时，表4.10和图4.17都可以论证得到其余六种治疗方式对最终水肿体积发展模式的影响较小，它们与发展模式之间为较低的弱相关联系。

4.5 子问题四：血肿体积、水肿体积及治疗方法的关系

1、血肿体积与水肿体积之间的关系

（1）血肿体积大小与水肿体积大小之间的相关性：

考虑到两者的关系并不是线性的，因此采用斯皮尔曼相关系数来进行计算，而非皮尔逊相关系数。最终得到两者体积之间的斯皮尔曼相关系数 $\rho = 0.588$，P 值 $P = 3.3\times10^{-5}$，显然小于显著性水平（通常为 0.05 或 0.01），因此可以得出结论，该相关性计算结果在统计上是显著的，而不仅仅是由于随机因素引起的。

（2）血肿体积变化量与水肿体积变化量之间的相关性：

仍然采用斯皮尔曼相关系数进行分析，得到两者体积变化量之间的相关系数为 $\rho = 0.53$，P 值 $P = 1.6\times10^{-8}$，所以该相关计算结果也是显著的。

表 4.11　血肿体积与水肿体积之间的关系

	血肿体积大小	血肿体积变化量
水肿体积大小	$\rho = 0.58$ $P = 3.3\times10^{-5}$	/
水肿体积变化量	/	$\rho = 0.53$ $P = 1.6\times10^{-8}$

因此，可以得出结论，血肿体积与水肿体积的相关性虽然并不属于是强相关，但是中等程度上正相关的。

2、血肿体积与治疗方案之间的关系

（1）血肿体积大小与治疗方案之间的联系：

由图4.18可以分析得出，采用脑室引流方案的患者，往往具有远超出平均水平的血肿体积，采取降颅压治疗的患者的血肿体积会略超出平均值，而是否采取其他五种治疗方案并没有与血肿体积大小有显著的联系。

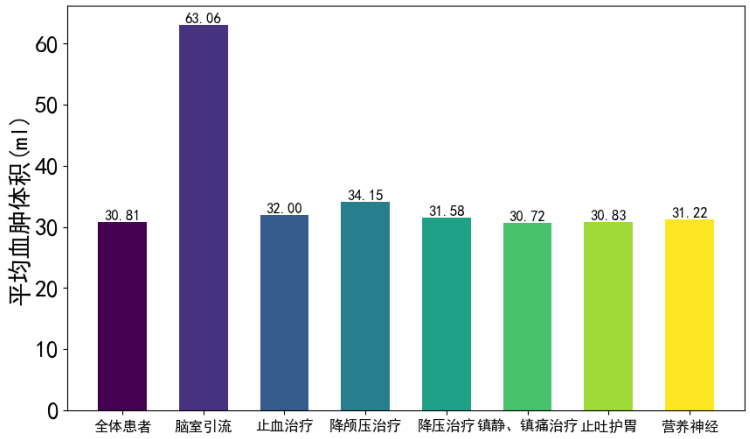


图 4.18　不同治疗方式对血肿体积的影响

接下来使用 Spearman 相关系数进行论证。平均血肿体积与治疗方案的相关系数表如下所示:

表 4.12　平均血肿体积与治疗方案的相关系数表

	水肿体积	脑室引流	止血治疗	降颅压治疗	降压治疗	镇静、镇痛治疗	止吐护胃	营养神经
血肿体积	0.58	0.33	0.18	0.27	0.1	-0.05	-0.08	0.06

从表中可以得到，水肿体积是与血肿体积大小关联性最大的因素，除此之外，治疗方案中与血肿体积关联最大的是脑室引流和降颅压治疗，这与图片中的结论相符合，往往采取脑室引流的患者血肿体积较大，采取降颅压治疗的患者次之，其余六个治疗方案则类似。

（2）血肿体积变化量与治疗方案之间的联系：

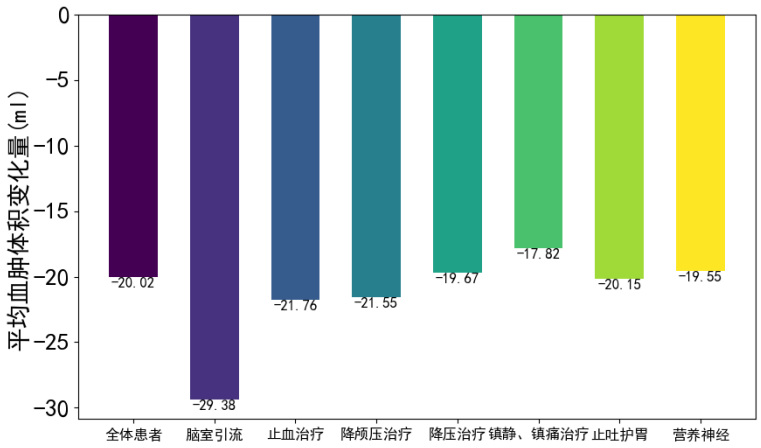


图 4.19　不同治疗方式对血肿体积变化量的影响

从图4.19中可以分析得到，在采取相应的治疗方案后，最终患者的血肿体积往往会相

较于发病时下降，同时采取脑室引流治疗方案的患者，血肿体积下降最显著，对下降血肿体积最有帮助。

平均血肿体积变化量与治疗方案的相关系数表如表4.13所示。

表 4.13　平均血肿体积变化量与治疗方案的相关系数表

	水肿体积变化量	脑室引流	止血治疗	降颅压治疗	降压治疗	镇静、镇痛治疗	止吐护胃	营养神经
血肿体积变化量	0.53	-0.02	-0.06	-0.19	0.01	0.05	-0.04	0.13

从表格中可以分析得到，水肿体积变化量仍是血肿体积变化量相关性最高的，但是脑室引流与血肿体积变化之间并没有呈较高的负相关之间的关系，这与图4.19中得出的结论相悖。

团队深入探究发现由于采取脑室引流的患者只有六位，该方案下患者样本数量较少，因此为更好的分析治疗方案帮助血肿体积减小或增加的实际情况，对所有治疗后血肿体积增加的患者标记为 1，血肿体积减小的患者标记为 0，可得到示意图4.20。

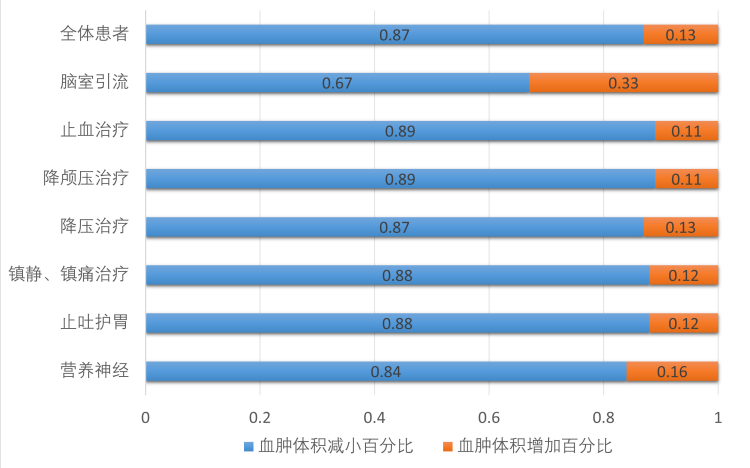


图 4.20　治疗方案中患者血肿体积增加、减少占比图

在上图中，我们可以得到所有治疗方案对于降低血肿体积都有一定的效果，止血治疗和降颅压治疗对于降低血肿体积是最好的。而虽然脑室引流在降低血肿体积的绝对值上最大的，但是采用脑室引流的患者中有 1/3 最终血肿体积增加，大于了其他六个方案。

因此对这六个患者数据分析后得到了脑室引流方案对不同人可以大幅降低血肿体积或者较高的概率增加血肿体积的矛盾特性，由于采用该方案的患者较少，这个结论可能是错误的，但在当前 100 个患者的样本中是正确的。

3、水肿体积与治疗方案之间的关系

（1）水肿体积大小与治疗方案之间的关系：

从图4.21中可以分析得到，采取脑室引流方案的患者往往具有较高的水肿体积，而是否采取其余六个方案与水肿体积大小并没有显著的联系。

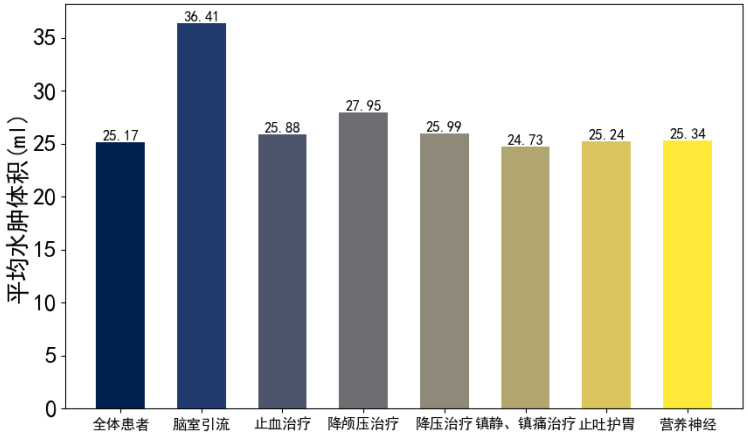


图 4.21　不同治疗方式对水肿体积的影响

平均水肿体积与治疗方案的相关系数表如表4.15所示。

表 4.14　平均水肿体积与治疗方案的相关系数表

	血肿体积	脑室引流	止血治疗	降颅压治疗	降压治疗	镇静、镇痛治疗	止吐护胃	营养神经
水肿体积	0.58	0.31	0.1	0.17	0.15	-0.08	0	0.04

从图4.21和表4.15可以分析得到，除了血肿体积外，脑室引流是与水肿体积关联度最高的治疗方案，采取脑室引流方案的患者往往具有较高的水肿体积，而是否采取其余六个方案与水肿体积大小并没有显著的联系。

（2）水肿体积变化量与治疗方案之间的关系：

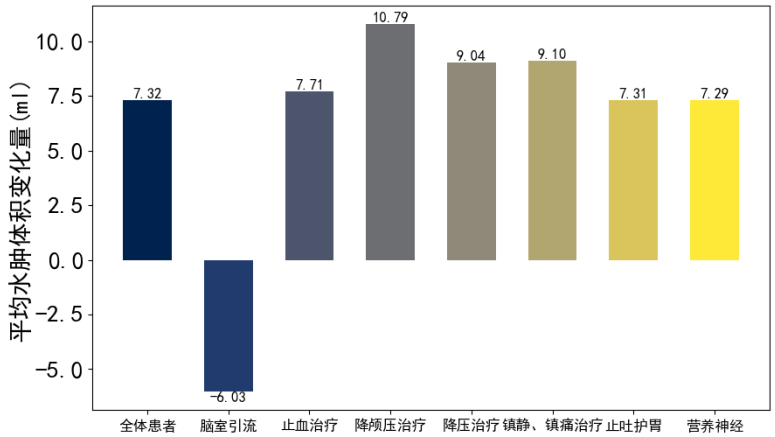


图 4.22　不同治疗方式对水肿体积变化量的影响

从图4.22中可以分析得到，采取相应的治疗方案后，所有患者的平均水肿体积会上升。但同时采取脑室引流方案的患者，是唯一平均水肿体积下降的。

平均水肿体积变化量与治疗方案的相关系数表如表4.15所示。

表 4.15　平均水肿体积与治疗方案的相关系数表

	血肿体积变化量	脑室引流	止血治疗	降颅压治疗	降压治疗	镇静、镇痛治疗	止吐护胃	营养神经
水肿体积变化量	0.53	-0.06	0	0.15	0.25	0.1	-0.02	0.01

从相关性分析表中可以得到降颅压治疗、降压治疗、镇静阵痛治疗都与水肿体积呈小幅度的正相关，与图中对应治疗方案平均水肿变化量大于全体患者平均值相对应。

考虑到不同患者下降/增加的幅度不同，因此仍然对所有治疗方案最终下降/增加水肿体积的患者比例进行分析，得到图4.23。

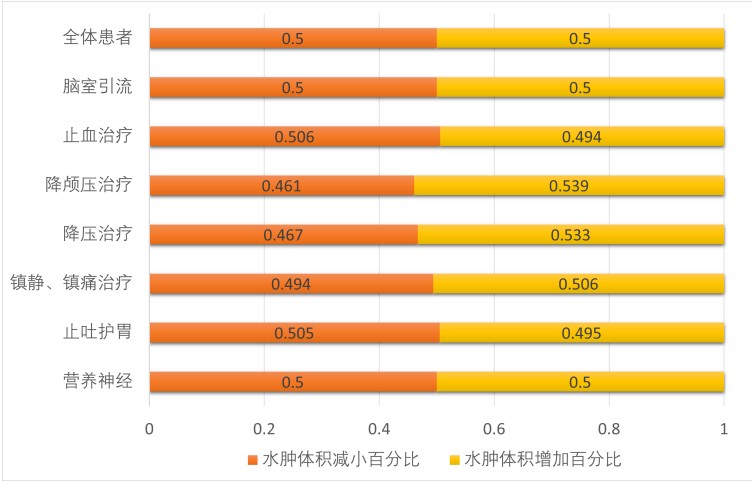


图 4.23　治疗方案中患者水肿体积增加、减少占比图

从图4.23我们可以得到，最终患者的水肿体积减小、增加的比例相同，说明只是全体患者中水肿体积增加的幅度大于减小幅度，导致全体患者水肿体积平均值计算是大于 0 的。

同时，对于脑室引流方案，可以进一步分析出虽然水肿体积平均量是下降的，但是增加/减小的患者数量是相同的。在水肿体积增加或减少的比例上，所有治疗方案并没有一个突出的表现。

5 问题三：预后预测建模及分析

5.1 问题分析

问题三旨在预测出血性脑卒中患者的预后，并探索可能影响预后的关键因素。具体来说，我们解决了以下三个子问题：

（1）基于患者首次影像信息构建 90 天 mRS 评分预测模型：

首先，我们考虑了前 100 位患者（sub001 至 sub100）的个人史、疾病史、发病相关因素以及首次影像结果。这些信息将被用来构建一个预测模型，以估计所有患者（sub001 至 sub160）在发病后 90 天内的 mRS 评分。mRS 评分是一个有序等级变量，代表了患者的神经功能和生活质量，因此对于临床决策具有重要价值。

（2）基于患者所有影像信息构建 90 天 mRS 评分预测模型：

接着，我们扩展了模型的输入数据，纳入了“表 2”和“表 3”中前 100 位患者的所有时间点的影像检查结果。这个更全面的模型将用于预测所有具有随访影像检查的患者（sub001 至 sub100,sub131 至 sub160）在发病后 90 天内的 mRS 评分。通过使用更多的信息，我们期望提高预测的准确性和可靠性，有助于更好地了解患者的预后情况。

（3）分析预后与关键因素的关联关系：

最后，我们对出血性脑卒中患者的预后进行深入分析，探索患者的个人史、疾病史、治疗方法以及影像特征等因素与 90 天 mRS 评分之间的关联关系，以提供临床相关的建议。这一分析有助于识别哪些因素可能对患者的康复和预后产生积极或负面影响，有助于医疗团队更好地制定个性化治疗方案。

综上，问题三旨在通过建立预测模型和深入研究关键因素，为出血性脑卒中患者的预后提供更准确的估计，并为临床决策提供有力支持。这有助于改善患者的治疗和康复过程，提高其生活质量。

5.2 子问题一：基于首次影像的 mRS 评分预测模型

5.2.1 数据预处理

1、数据拼接

对于该问题将“表 1”中包含的患者信息与“表 2”和“表 3”中患者首次影像检查结果进行了拼接，以整合来自不同数据表的信息。该过程示意图如图5.24所示。

2、独热编码

对于“表 1”中的分类特征，如个人史、疾病史、发病相关信息等，我们采用了独热编码的方法进行处理。

3、主成分分析

对于“表 2”和“表 3”中的影像学特征数据，我们进行了标准化处理，使得所有特征

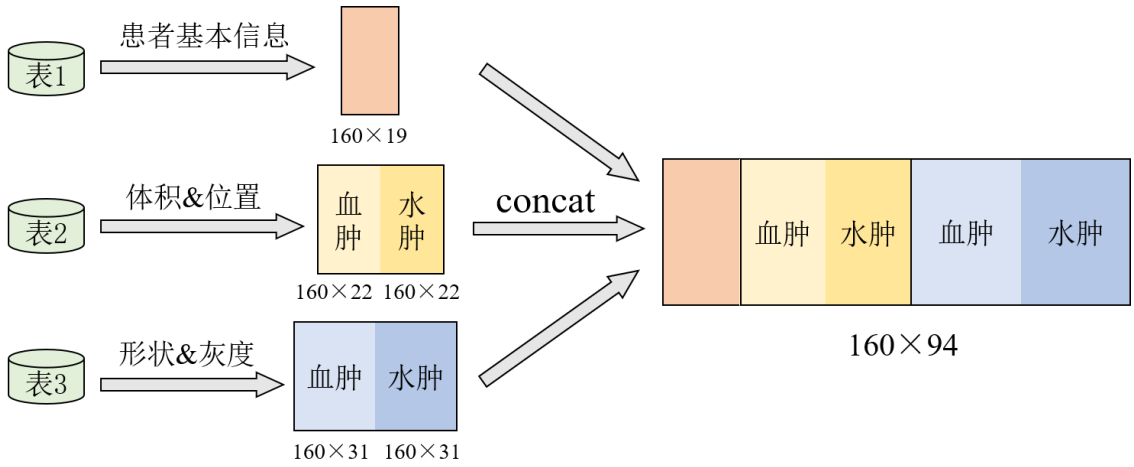


图 5.24　数据拼接

都具有相同的尺度。此外，由于影像学信息包含了大量的特征，可能存在冗余信息或高度相关的特征。为了降低维度并消除冗余，我们采用了 PCA 方法减少数据的维度，同时保留尽可能多的信息。

5.2.2 模型构建

在该问题中，我们选择了随机森林和 XGBoost 算法作为我们的主要建模工具。同时，使用了 GridSearch 方法选取模型的最佳超参数，包括学习率、树的深度、树的数量等。

5.2.3 结果分析

分别使用两种模型对数据进行训练，最终得到的模型准确率如表5.16所示。

表 5.16　随机森林和 XGBoost 模型准确率对比

模型	测试集上的准确率
随机森林	20%
XGBoost	35%

由表可知，XGBoost 模型的准确率更高。因此，对于本题我们使用 XGBoost 模型对患者（sub001 至 sub160）90 天 mRS 评分进行预测，并将预测结果填写在“表 4”I 字段（预测 mRS（基于首次影像））中。同时也可见本文的附录 C。

5.3 子问题二：基于所有影像的 mRS 评分预测模型

与子问题一不同，该问题考虑了患者所有的影像检查信息，以提供更全面的预测。考虑到数据包含时间序列信息，我们采用了两种不同的机器学习模型：XGBoost 和循环神经网络（RNN）。

5.3.1 XGBoost 模型

1、数据预处理

对于该问题，我们首先对数据进行二维拼接，即类似于子问题一中的处理，将“表 1”中包含的患者信息与“表 2”和“表 3”中的每次影像检查时间点及对应检查结果进行拼接，最终将患者的个人信息与多次影像检查结果按时间序列整合到一张表中。

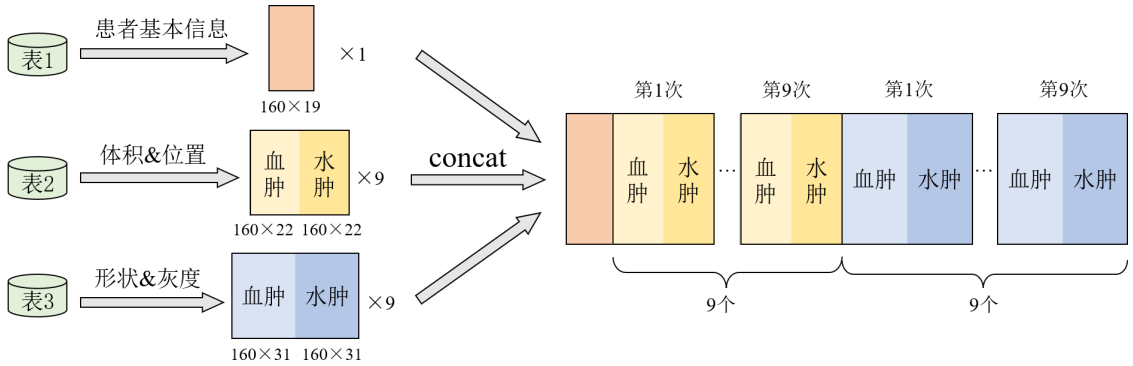


图 5.25　二维拼接

由于每位患者进行影像检查的次数不同，即数据中存在缺失值。对于 XGBoost 算法，它可以自行处理缺失值而无需我们进行额外处理。

2、模型训练

我们使用了 GridSearch 方法选取模型的最佳超参数，其中几个主要参数如表5.17所示。

表 5.17　XGBoost 参数表格

eta	0.2154
n_estimators	100
gamma	1
max_depth	2
min_child_weight	2
reg_lambda	50
reg_alpha	0

5.3.2 RNN 模型

1、数据预处理

对于 RNN 模型，我们选择对数据进行三维拼接，这一方法更加注重时序信息的处理。我们将每个患者的影像检查次数单独作为一维数据，然后在每个维度上进行“表 1”中的

患者信息和“表 2”、“表 3”中的影像检查结果的数据拼接。这种方法允许我们更好地考虑时间维度对预后情况的影响。

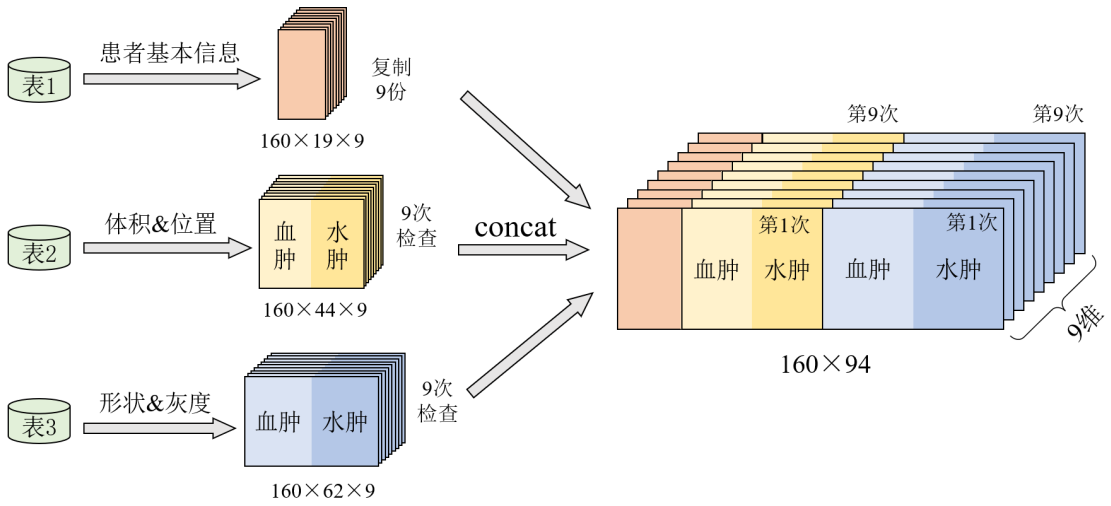


图 5.26　三维拼接

对于拼接后的缺失值处理，我们在问题一中已经测试得出填充“0”的准确率最高，因此这里沿用该方法，对缺失的数据填充“0”。

2、模型构建

对于本题，我们的模型结构包括以下几个关键部分：

（1）RNN 层：我们首先使用 RNN 层来处理序列数据。

循环神经网络（Recurrent Neural Network，简称 RNN）[8] 是一类在序列数据上表现出色的深度学习模型。与传统神经网络不同，RNN 具有一种递归的结构，能够有效地处理具有时序性和序列依赖关系的数据。RNN 的核心结构包括一个循环单元（Recurrent Unit），通常表示为一个带有输入、输出和隐藏状态的神经元。隐藏状态可以理解为网络在处理序列数据时的内部记忆，它保留了之前处理过的信息，以帮助网络更好地理解当前输入。

RNN 使用反向传播算法来更新权重并学习数据的表示。然而，传统 RNN 在处理长序列时存在梯度消失或梯度爆炸的问题，这导致难以捕捉长距离依赖关系。为了解决这个问题，出现了一些改进的 RNN 变体，如长短时记忆网络（LSTM）和门控循环单元（GRU），它们更好地处理了梯度问题。

（2）全连接层（输出层）：我们在 RNN 的基础上引入了一个全连接层，也是输出层。其神经元数量等于本题多分类任务的类别数量，即 7。采用 softmax 激活函数，将模型的输出映射到一个概率分布上，表示每个类别的概率。

3、模型训练

对于该模型，我们使用的参数如表5.18所示。

表 5.18　RNN 参数表格

输入维度	104
隐藏层数量	64
输出维度	7
序列长度	9
lr	0.0001
epoch	5000
batch_size	20
损失函数	交叉熵损失
优化器	Adam

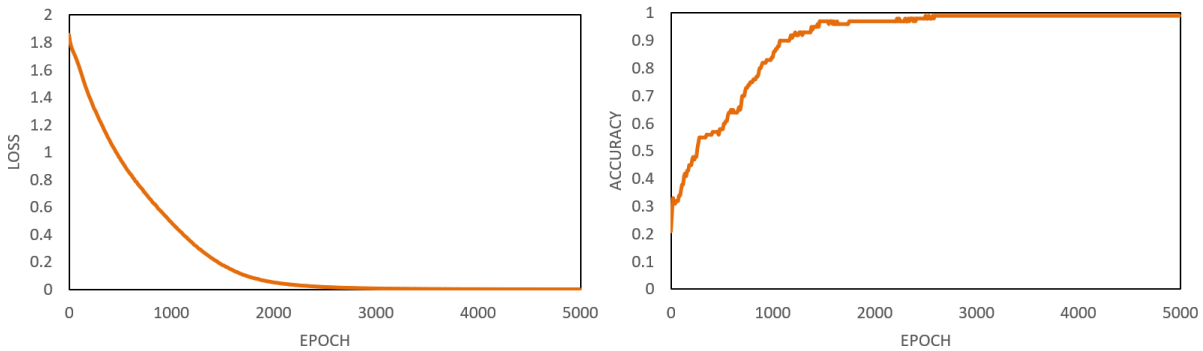
5.3.3 模型对比

分别使用 XGBoost 和 RNN 对数据进行训练和测试，最终得到的模型准确率如表5.19所示。

表 5.19　模型准确率对比

模型	测试集上的准确率
XGBoost	50%
RNN	98%

其中，RNN 模型在训练过程中的损失变化曲线和在测试集上的准确率变化曲线如图5.27所示。



(a) 损失变化曲线　　(b) 准确率变化曲线

图 5.27　RNN 训练结果

可以看到，相比于只考虑患者首次影像检查信息，当考虑到患者全部影像检查信息时

模型预测的准确性有大幅提升，这是因为将所有影像检查信息整合到模型中，可以综合考虑多个时间点的信息，从而减少了可能的信息丢失。这种综合信息的优势使得模型能够更全面地评估患者的情况，更准确地进行预测。

此外，RNN 模型的预测准确率远远大于 XGBoost 模型，说明了除更多特征引入之外，时序信息对于模型的重要性。这些时序信息包含了患者病情的演变趋势，而 RNN 能够有效地捕捉到这些信息之间的关联和变化趋势。

因此，我们使用 RNN 模型对所有含随访影像检查的患者（sub001 至 sub100, sub131 至 sub160）90 天 mRS 评分进行预测，预测结果填写在“表 4”的 J 字段中。同时也可见本文的附录 C。

5.4 子问题三：分析预后与关键因素的关联

对于该问题，我们将所有关键因素分为三个方面：个人史和疾病史相关因素、治疗方法相关因素、影像特征相关因素。统计分析这些因素与患者预后（90 天 mRS 评分）之间的关联关系，进而为临床相关决策提出建议。

5.4.1 个人史和疾病史相关因素

患者的个人史和疾病史等相关因素包括患者的年龄、性别、脑出血前 mRS 评分、高血压病史、卒中病史、糖尿病史、房颤史、冠心病史、吸烟史、饮酒史、血压 11 个因素。我们使用 Spearman 系数计算了这些因素与患者预后之间的相关性，结果如表5.20所示。

表 5.20　预后与个人史和疾病史相关性

年龄	性别	脑出血前 mRS 评分	高血压病史	卒中病史	糖尿病史	房颤史	冠心病史	吸烟史	饮酒史	血压
0.162	-0.067	-0.017	0.073	0.052	0.299	0.168	0.292	-0.220	-0.268	-0.076

此外，我们将相关性结果取绝对值后，绘制成了一张雷达图，如图5.28(a) 所示，用以直观展示不同因素对 90 天 mRS 评分的影响。

5.4.2 治疗方法相关因素

我们认为，手术治疗、药物治疗以及其他可能的治疗方法对患者预后也会产生不同影响。本题中包括脑室引流、止血治疗、降颅压治疗、降压治疗、镇静或镇痛治疗、止吐护胃、营养神经 7 个不同因素。这些因素与患者预后之间的相关性结果如表5.21所示。

预后与治疗方法相关性结果雷达图如图5.28(b) 所示。

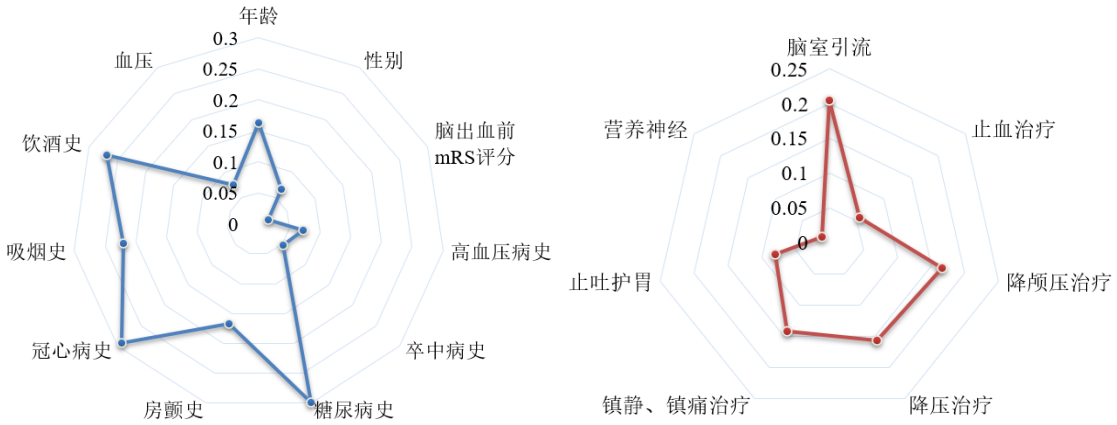
表 5.21　预后与治疗方法相关性

脑室引流	止血治疗	降颅压治疗	降压治疗	镇静、镇痛治疗	止吐护胃	营养神经
0.20502	-0.0561	0.165863	0.156147	0.142265	-0.08037	0.013455

5.4.3 影像特征相关因素

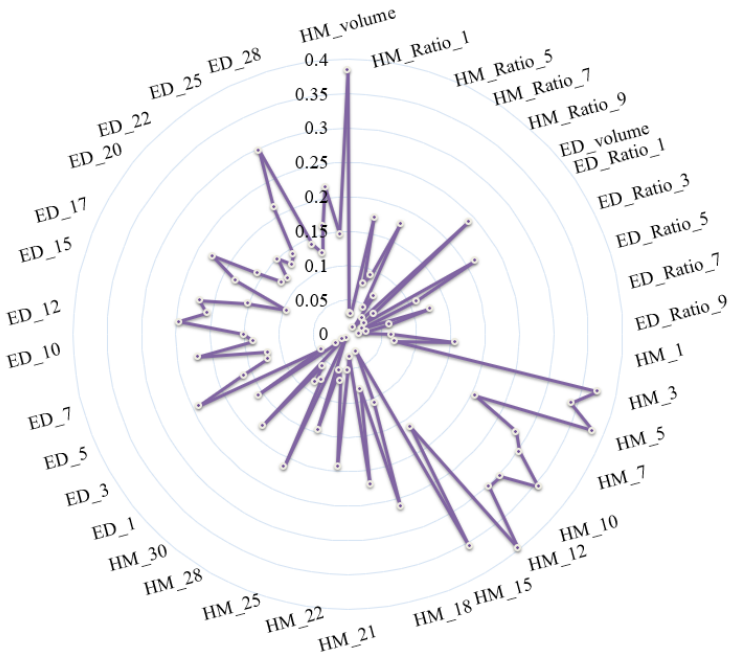
影像特征是关于出血性脑卒中患者病变的关键信息，包括血肿和水肿的体积、位置分布、形状特征和灰度分布等。预后与影像特征相关性结果雷达图如图5.28(c) 所示。

其中，HM_volume 和 ED_volume 分别对应“表 2”中患者血肿体积和水肿体积特



(a) 个人史和疾病史相关因素

(b) 治疗方法相关因素



(c) 影像特征相关因素

图 5.28　预后与关键因素相关性分析雷达图

征，HM_Ratio_1 至 HM_Ratio_31 分别表示“表 2”中患者血肿和水肿发生的十个不同位置，HM_1 至 HM_31 和 ED_1 至 ED_31 分别对应“表 3”中血肿和水肿的 31 个影像特征（包括形状特征和灰度特征）。

5.4.4 结果分析与临床建议

从图5.28(a) 中我们可以发现，个人史和疾病史中的糖尿病史以及冠心病史与出血性脑卒中患者的预后之间存在较大的相关性，并且由表5.20可知，均与患者预后成正相关，说明有相关病史的患者 90 天 mRS 评分越高，即患者的功能状态越差，残疾程度越高。

从图5.28(b) 中我们可以发现，脑室引流的治疗方法与患者预后相关性较大，并且由表5.21可知它们成正相关，而营养神经的治疗方法相关性最小。

从图5.28(c) 中我们可以发现，相比于水肿，血肿与患者预后相关性更大。其中，血肿体积，血肿形状特征中的 LeastAxisLength、Maximum2DDiameterColumn、SurfaceArea，以及血肿灰度特征中的 10Percentile 这几个因素相关性较大。

基于上述分析，我们为临床相关决策提供了一些建议：

（1）加强早期干预和管理：对于脑卒中患者，特别是那些具有糖尿病或冠心病病史，以及血肿体积较高的患者，早期干预和紧密监测尤为重要。医疗团队应该积极介入，并确保这些患者接受及时和有效的治疗，以减轻脑卒中的影响。

（2）个性化治疗计划：不同患者可能需要不同的治疗方法。根据患者的具体情况和病史，医疗团队应该制定个体化的治疗计划，并且尽量采取营养神经的治疗方法，可提供适当的营养支持和促进神经功能恢复，维持免疫功能、促进伤口愈合和预防并发症的发生。

（3）定期随访和监测：在临床实践中，患者应该接受定期的随访和监测，以确保治疗的有效性，并在需要时进行调整。医生应仔细记录患者的血肿特征，包括体积、形状、灰度分布等。医疗团队可以将患者的部分特征与预后进行关联，例如较大的血肿可能导致更差的预后，医生可以根据这些特征更准确地预测患者的康复潜力。

（4）多学科协作和创新：脑卒中治疗通常需要多学科的协作。医疗团队应该共同评估患者情况，以确定最合适的治疗方法。同时，进一步的研究和创新可以探索不同治疗方法的长期影响和效果。通过持续的科研工作，可以不断改进脑卒中治疗策略，提高患者的生活质量和预后。

然而我们发现，大多数特征的相关系数都较低。此外，吸烟史、饮酒史与患者预后呈负相关关系，但是大量研究和医学证据表明，吸烟和饮酒通常被认为是不健康的生活方式，会造成脑卒中的发病风险增加。因此我们推断出现该情况的原因是本题的数据量不足，且个体的健康状况和脑卒中的严重程度会因多种因素而异，因此在制定治疗和康复计划时，应该综合考虑患者的具体情况，包括吸烟和饮酒的习惯，以制定最合适的治疗策略。

总之，脑卒中是一种复杂的疾病，对患者和医疗团队都提出了挑战。综合多种治疗和管理策略，个体化治疗计划以及多学科交叉创新，相信脑卒中患者的预后可以得到改善。

6 模型总结

6.1 模型的优点

1. 本文充分考虑了数据的特征重要性，在输入预测模型前使用标准化、Spearman 相关性分析以及 PCA 等数据预处理方法；

2. 本文在对问题分析时考虑较为全面，采用多个模型或算法进行求解，对一个模型进行纵向最佳超参选取，横向模型对比；

3. 对医学上的数据进行绝对大小和变化量两个维度分析，保证了最终结果的稳健性和完善性；

4. 充分融合机器学习以及可视化的统计分析的优点，如在问题二中治疗方法与水肿、血肿体积的关系分析，正视脑室引流患者较少的问题，给出“矛盾性”的原因以及更为深入的结果答案；

5. 不仅能够有效地处理高维数据和非线性关系，同时也考虑到了对具有时序关系的数据的处理。

6.2 模型的不足

1. 模型对不同患者间数据的独立性分布要求较高，并要求同一位患者的所有随访数据为存在潜在强相关关系；

2. 部分模型如 XGBoost 和 RNN 等模型的预测性能很好，但它们的解释性相对较差。若要在医学领域进行预测分析，需要改进并选择相应的模型解释工具。

6.3 未来的改进方向

1. 获取更多临床真实数据，使用大规模数据集进行训练和测试，提高模型的鲁棒性：

2. 可以考虑使用模型融合技术，将多个模型的预测结果进行组合，以进一步提高预测性能。

参考文献

[1] Nonparametric Statistical Methods, New York, NY: Springer New York, 121-162, 2009.

[2] Kurita T, Principal Component Analysis (PCA), Computer Vision, A Reference Guide, 2014.

[3] Cramer J S, The Origins of Logistic Regression, Econometrics eJournal, 2002.

[4] Ho T K, Random decision forests, Proceedings of 3rd International Conference on Document Analysis and Recognition: volume 1, 278-282 vol.1, 1995.

[5] Hunt E B, Marin J, Stone P J, Experiments in induction, 1966.

[6] Chen T, Guestrin C, XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[7] Refaeilzadeh P, Tang L, Liu H, Cross-Validation, Boston, MA: Springer US, 532-538, 2009.

[8] Zaremba W, Sutskever I, Vinyals O, Recurrent Neural Network Regularization, ArXiv, abs/1409.2329, 2014.

附录 A　问题一结果表格

最终得到的问题一的相关结果如下表所示，我们填写了以下三列数据：

（1）在列“是否发生血肿扩张”中，我们判断了前 100 个患者（sub001 至 sub100）发病后 48 小时内是否发生血肿扩张事件，“1”代表“是”，“0”代表“否”。

（2）在列“血肿扩张时间”中，对于被我们判断为在发病后 48 小时内发生了发生血肿扩张事件的患者，我们填写了其对应的血肿扩张时间，单位为小时。对于被我们判断为没有发生血肿扩张的患者，该列数据不填。

（3）在列“血肿扩张预测概率”中，我们记录了对所有患者（sub001 至 sub160）发生血肿扩张概率的预测结果，取值范围 0-1，且小数点后保留 4 位数。

	首次影像检查流水号	问题 1：血肿扩张		
		是否发生血肿扩张 1 是，0 否	血肿扩张时间 单位：小时	血肿扩张预测概率
sub001	20161212002136	0		0.1671
sub002	20160406002131	0		0.0025
sub003	20160413000006	1	9.52	0.2859
sub004	20161215001667	0		0.0274
sub005	20161222000978	1	26.47	0.1403
sub006	20161110001074	0		0.0157
sub007	20161208000139	0		0.0018
sub008	20161219000091	0		0.0804
sub009	20161031001987	1	40.06	0.8367
sub010	20161012002008	0		0.0448
sub011	20160209000219	0		0.3710
sub012	20161031001142	0		0.0044
sub013	20161124000397	0		0.0514
sub014	20160513001799	0		0.0382
sub015	20161013001234	0		0.0150
sub016	20161130000004	0		0.0294
sub017	20160510002436	1	14.87	0.8938
sub018	20160602001707	0		0.0499
sub019	20160117000135	0		0.0126
sub020	20160723000013	0		0.0097
sub021	20160317001244	0		0.2139

sub022	20160803001239	0		0.0085
sub023	20160321000142	0		0.0076
sub024	20170802000637	0		0.1738
sub025	20171226002293	0		0.0609
sub026	20171008000512	0		0.0239
sub027	20170206000071	0		0.2264
sub028	20171013002097	0		0.3238
sub029	20170607000010	0		0.0104
sub030	20171025000480	0		0.0166
sub031	20170307002130	0		0.5801
sub032	20171009000137	0		0.0248
sub033	20170115000362	1	30.81	0.0200
sub034	20170119000729	0		0.0514
sub035	20171014001244	0		0.0072
sub036	20170204001714	1	39.50	0.0555
sub037	20170426000005	0		0.3901
sub038	20170518002194	1	15.81	0.9139
sub039	20170425002487	1	29.18	0.8469
sub040	20170902000876	0		0.0501
sub041	20171002000282	0		0.7112
sub042	20170420000636	0		0.0554
sub043	20170325000428	0		0.0370
sub044	20170528000084	0		0.0199
sub045	20170324001892	0		0.0550
sub046	20170511000016	0		0.2663
sub047	20171019001652	0		0.0497
sub048	20170402000556	1	12.86	0.2924
sub049	20171005000770	0		0.0134
sub050	20171105000372	0		0.0100
sub051	20170422000935	0		0.0060
sub052	20170608001310	0		0.0523
sub053	20170511001392	0		0.0813
sub054	20170612002216	1	16.23	0.5747

sub055	20170316001977	0		0.0678
sub056	20170120000152	0		0.0298
sub057	20170825001844	1	14.62	0.3517
sub058	20170125000984	0		0.0706
sub059	20170912002314	0		0.4945
sub060	20180109000613	1	23.73	0.1092
sub061	20180226000725	1	6.54	0.4454
sub062	20181221002264	0		0.1822
sub063	20181020001229	0		0.1180
sub064	20180801000501	0		0.0053
sub065	20180131001727	0		0.0861
sub066	20181208000909	0		0.0179
sub067	20181207001317	0		0.0365
sub068	20180412001426	0		0.0389
sub069	20180619001505	0		0.0483
sub070	20180427000292	1	9.65	0.0263
sub071	20181103001264	0		0.0321
sub072	20181007000826	0		0.0559
sub073	20180911001645	0		0.0372
sub074	20180719000020	0		0.6245
sub075	20180428001767	0		0.0077
sub076	20180619002401	1	15.99	0.7453
sub077	20180503002304	1	14.12	0.2709
sub078	20180929000040	0		0.7971
sub079	20180929000037	1	27.85	0.7551
sub080	20180130001917	1	20.57	0.0934
sub081	20180120000249	1	27.42	0.0837
sub082	20180221000793	0		0.0164
sub083	20181004000706	0		0.0197
sub084	20180716000006	0		0.0283
sub085	20181127002511	0		0.3470
sub086	20180108000002	0		0.0049
sub087	20180216000198	0		0.0182

sub088	20180521000314	0		0.0373
sub089	20180314002318	0		0.0142
sub090	20180910002366	0		0.0117
sub091	20181019001130	0		0.0129
sub092	20181116001089	1	11.92	0.2834
sub093	20181214000208	0		0.0674
sub094	20180412001795	0		0.0337
sub095	20180316001329	1	7.43	0.2199
sub096	20180802001789	0		0.0363
sub097	20181010000767	0		0.0184
sub098	20180612002507	1	42.76	0.7673
sub099	20180620002296	1	17.67	0.3049
sub100	20180314000010	0		0.0391
sub101	20180311000432			0.0427
sub102	20180708000024			0.5335
sub103	20181015001677			0.0874
sub104	20190105000694			0.0072
sub105	20190108002459			0.0275
sub106	20190519000853			0.0059
sub107	20190526000209			0.0024
sub108	20190701002502			0.0096
sub109	20190716000013			0.3373
sub110	20190717001385			0.0242
sub111	20190727000556			0.0056
sub112	20190803000014			0.0066
sub113	20190901000442			0.0052
sub114	20190903000373			0.1957
sub115	20190904002299			0.0102
sub116	20190906001283			0.0080
sub117	20190917002094			0.0517
sub118	20190923002580			0.1671
sub119	20191003000352			0.0026
sub120	20191004001025			0.0069

sub121	20191027000468			0.0016
sub122	20191028002738			0.0358
sub123	20191024001280			0.0015
sub124	20191030001526			0.0143
sub125	20191111002510			0.0021
sub126	20191126002590			0.0449
sub127	20191127002176			0.3277
sub128	20191208000592			0.2206
sub129	20191224000008			0.0015
sub130	20191228000237			0.6161
sub131	20160413000006			0.1204
sub132	20161215001667			0.7099
sub133	20200112000228			0.0021
sub134	20200101000392			0.0097
sub135	20201130003288			0.0190
sub136	20201203002778			0.0267
sub137	20201217002368			0.0067
sub138	20200403000012			0.0549
sub139	20200814000015			0.6813
sub140	20200412000331			0.0218
sub141	20200711001264			0.0103
sub142	20200411000014			0.0093
sub143	20200214000572			0.0523
sub144	20200124000041			0.0414
sub145	20200613001086			0.0115
sub146	20200531000253			0.0135
sub147	20201220000155			0.0520
sub148	20200609001016			0.0058
sub149	20200409001101			0.0083
sub150	20200118000372			0.0010
sub151	20201023001238			0.7569
sub152	20201109000009			0.0107
sub153	20201212001420			0.0020

sub154	20201129000299			0.1377
sub155	20200301000025			0.0095
sub156	20200306000927			0.0568
sub157	20201009003102			0.0868
sub158	20200410001952			0.1899
sub159	20200218000582			0.1402
sub160	20200821002584			0.1061

附录 B　问题二结果表格

最终得到的问题二的相关结果如下表所示，我们填写了以下三列数据：

（1）在列“残差（全体）”中，我们记录了前 100 个患者（sub001 至 sub100）水肿体积真实值和所拟合曲线之间存在的残差，单位为 ml。

（2）在列“残差（亚组）”中，我们记录了前 100 个患者（sub001 至 sub100）水肿体积真实值和所在亚组所拟合的曲线之间存在的残差，单位为 ml。

（3）在列“所属亚组”中，我们记录了前 100 个患者（sub001 至 sub100）经过我们分类后所在的亚组号，取值为 1、2、3 或 4，分别代表亚组 1、亚组 2、亚组 3、亚组 4。

	首次影像检查流水号	问题 2：水肿体积进展曲线		
		残差（全体）	残差（亚组）	所属亚组
sub001	20161212002136	63.7035	51.8307	3
sub002	20160406002131	17.0818	12.2093	1
sub003	20160413000006	24.3250	7.3770	4
sub004	20161215001667	13.2583	11.8217	2
sub005	20161222000978	17.8951	8.3312	2
sub006	20161110001074	53.7248	44.4881	2
sub007	20161208000139	35.1463	26.2848	1
sub008	20161219000091	11.6836	8.0092	4
sub009	20161031001987	9.2700	6.6324	3
sub010	20161012002008	16.0337	14.3033	3
sub011	20160209000219	15.3550	5.7098	1
sub012	20161031001142	18.9345	9.5472	1
sub013	20161124000397	9.1941	2.1939	1
sub014	20160513001799	10.8572	7.4571	3
sub015	20161013001234	33.8259	24.7132	3
sub016	20161130000004	33.2405	22.6777	4
sub017	20160510002436	8.9880	6.4378	3
sub018	20160602001707	22.3573	13.2318	3
sub019	20160117000135	13.0731	10.6560	3
sub020	20160723000013	12.0761	6.4380	1
sub021	20160317001244	14.0875	5.9417	1
sub022	20160803001239	10.9998	1.5003	1
sub023	20160321000142	24.9939	20.5877	1

sub024	20170802000637	15.2820	6.0560	1
sub025	20171226002293	12.2013	4.3879	3
sub026	20171008000512	15.6738	6.3543	3
sub027	20170206000071	19.2853	18.8161	3
sub028	20171013002097	18.1873	10.4074	3
sub029	20170607000010	22.8400	20.2473	3
sub030	20171025000480	45.1294	31.5801	3
sub031	20170307002130	18.3230	4.8561	4
sub032	20171009000137	19.0042	12.2807	1
sub033	20170115000362	11.2519	3.9261	1
sub034	20170119000729	31.7161	10.3650	4
sub035	20171014001244	7.5303	3.4301	1
sub036	20170204001714	18.5072	12.3821	1
sub037	20170426000005	21.4142	13.5221	3
sub038	20170518002194	16.7117	11.9948	3
sub039	20170425002487	38.0649	28.1844	3
sub040	20170902000876	23.2421	14.3394	3
sub041	20171002000282	19.1339	17.9513	3
sub042	20170420000636	7.9777	4.0384	1
sub043	20170325000428	15.4962	8.3385	1
sub044	20170528000084	26.2698	17.3705	3
sub045	20170324001892	9.5895	9.1587	3
sub046	20170511000016	41.5903	32.3107	1
sub047	20171019001652	17.7814	12.3511	3
sub048	20170402000556	19.9634	7.3599	4
sub049	20171005000770	20.6265	10.6197	1
sub050	20171105000372	8.3719	5.5345	1
sub051	20170422000935	5.6737	4.5786	1
sub052	20170608001310	22.7811	13.4540	3
sub053	20170511001392	8.0246	6.5033	3
sub054	20170612002216	12.4361	12.2880	3
sub055	20170316001977	11.2059	4.4663	3
sub056	20170120000152	14.6636	11.8261	1

sub057	20170825001844	8.2148	0.4550	1
sub058	20170125000984	14.3875	6.0873	3
sub059	20170912002314	19.7699	11.0122	1
sub060	20180109000613	20.1298	20.0463	1
sub061	20180226000725	39.4651	20.9888	4
sub062	20181221002264	14.4997	10.8193	3
sub063	20181020001229	33.3331	23.1564	3
sub064	20180801000501	14.0383	13.4820	2
sub065	20180131001727	26.0208	19.1411	2
sub066	20181208000909	21.1305	17.9321	4
sub067	20181207001317	14.4284	8.4099	1
sub068	20180412001426	22.4267	17.8746	1
sub069	20180619001505	21.1598	17.2387	1
sub070	20180427000292	15.9030	12.7611	3
sub071	20181103001264	32.1221	26.6704	1
sub072	20181007000826	14.3958	11.6706	3
sub073	20180911001645	14.9933	9.1980	1
sub074	20180719000020	40.3584	32.2931	3
sub075	20180428001767	15.8733	11.5560	1
sub076	20180619002401	10.4469	3.0561	1
sub077	20180503002304	30.9209	23.6564	4
sub078	20180929000040	19.5510	11.8120	1
sub079	20180929000037	13.0366	4.3017	1
sub080	20180130001917	8.7732	3.0577	1
sub081	20180120000249	34.4234	32.7701	3
sub082	20180221000793	10.1099	5.2058	2
sub083	20181004000706	13.7807	9.2278	3
sub084	20180716000006	14.8274	6.6861	3
sub085	20181127002511	26.4521	16.2083	2
sub086	20180108000002	17.6346	9.1064	1
sub087	20180216000198	13.1041	5.5531	1
sub088	20180521000314	6.1124	5.2473	2
sub089	20180314002318	12.8765	9.2173	3

sub090	20180910002366	38.2127	31.0318	4
sub091	20181019001130	29.3782	20.4934	3
sub092	20181116001089	14.5312	6.0058	3
sub093	20181214000208	13.8834	6.6288	1
sub094	20180412001795	8.2386	2.4375	1
sub095	20180316001329	78.2321	43.5361	4
sub096	20180802001789	14.5083	4.5942	1
sub097	20181010000767	14.5884	8.5691	1
sub098	20180612002507	23.8681	19.6950	1
sub099	20180620002296	22.4396	16.8948	4
sub100	20180314000010	16.2261	8.2999	1

附录 C　问题三结果表格

最终得到的问题三的相关结果如下表所示，我们填写了以下两列数据：

（1）在列“预测 mRS（基于首次影像）”中，我们记录了使用基于首次影像结果得到的预测模型，对所有患者（sub001 至 sub160）90 天 mRS 评分的预测结果，取值范围为 0,1,2,3,4,5,6。

（2）在列“预测 mRS”中，我们记录了使用基于所有影像结果得到的预测模型，对所有含随访影像检查的患者（sub001 至 sub100,sub131 至 sub160）90 天 mRS 评分的预测结果，取值范围为 0,1,2,3,4,5,6。不含随访影像检查的患者（sub131 至 sub160）所在行不填。

	首次影像检查流水号	问题 3：预后预测建模	
		预测 mRS（基于首次影像）	预测 mRS
sub001	20161212002136	0	4
sub002	20160406002131	0	0
sub003	20160413000006	5	5
sub004	20161215001667	4	4
sub005	20161222000978	3	3
sub006	20161110001074	5	5
sub007	20161208000139	2	2
sub008	20161219000091	4	4
sub009	20161031001987	3	3
sub010	20161012002008	3	3
sub011	20160209000219	2	1
sub012	20161031001142	0	0
sub013	20161124000397	1	1
sub014	20160513001799	2	2
sub015	20161013001234	2	2
sub016	20161130000004	5	5
sub017	20160510002436	3	3
sub018	20160602001707	0	0
sub019	20160117000135	4	1
sub020	20160723000013	2	2
sub021	20160317001244	3	3
sub022	20160803001239	2	2
sub023	20160321000142	0	1

sub024	20170802000637	1	1
sub025	20171226002293	2	2
sub026	20171008000512	3	3
sub027	20170206000071	6	6
sub028	20171013002097	4	4
sub029	20170607000010	4	4
sub030	20171025000480	5	5
sub031	20170307002130	2	4
sub032	20171009000137	2	0
sub033	20170115000362	5	5
sub034	20170119000729	2	4
sub035	20171014001244	5	5
sub036	20170204001714	3	3
sub037	20170426000005	3	3
sub038	20170518002194	2	2
sub039	20170425002487	1	1
sub040	20170902000876	1	1
sub041	20171002000282	3	3
sub042	20170420000636	0	0
sub043	20170325000428	5	5
sub044	20170528000084	3	3
sub045	20170324001892	2	1
sub046	20170511000016	0	2
sub047	20171019001652	1	2
sub048	20170402000556	3	3
sub049	20171005000770	1	1
sub050	20171105000372	1	1
sub051	20170422000935	0	0
sub052	20170608001310	3	3
sub053	20170511001392	2	2
sub054	20170612002216	4	3
sub055	20170316001977	4	4
sub056	20170120000152	0	4

sub057	20170825001844	3	3
sub058	20170125000984	5	5
sub059	20170912002314	0	0
sub060	20180109000613	4	4
sub061	20180226000725	4	4
sub062	20181221002264	2	2
sub063	20181020001229	2	2
sub064	20180801000501	3	3
sub065	20180131001727	3	3
sub066	20181208000909	6	6
sub067	20181207001317	2	2
sub068	20180412001426	2	2
sub069	20180619001505	6	6
sub070	20180427000292	3	3
sub071	20181103001264	4	1
sub072	20181007000826	5	5
sub073	20180911001645	1	1
sub074	20180719000020	1	0
sub075	20180428001767	3	3
sub076	20180619002401	5	5
sub077	20180503002304	0	6
sub078	20180929000040	1	1
sub079	20180929000037	1	1
sub080	20180130001917	1	1
sub081	20180120000249	3	3
sub082	20180221000793	2	2
sub083	20181004000706	5	5
sub084	20180716000006	2	1
sub085	20181127002511	2	2
sub086	20180108000002	0	0
sub087	20180216000198	1	1
sub088	20180521000314	2	2
sub089	20180314002318	1	1

sub090	20180910002366	5	5
sub091	20181019001130	2	2
sub092	20181116001089	5	5
sub093	20181214000208	2	2
sub094	20180412001795	5	5
sub095	20180316001329	4	4
sub096	20180802001789	4	4
sub097	20181010000767	2	2
sub098	20180612002507	5	5
sub099	20180620002296	3	3
sub100	20180314000010	2	2
sub101	20180311000432	5	
sub102	20180708000024	2	
sub103	20181015001677	1	
sub104	20190105000694	0	
sub105	20190108002459	2	
sub106	20190519000853	5	
sub107	20190526000209	0	
sub108	20190701002502	2	
sub109	20190716000013	2	
sub110	20190717001385	2	
sub111	20190727000556	5	
sub112	20190803000014	4	
sub113	20190901000442	4	
sub114	20190903000373	5	
sub115	20190904002299	5	
sub116	20190906001283	4	
sub117	20190917002094	2	
sub118	20190923002580	6	
sub119	20191003000352	2	
sub120	20191004001025	1	
sub121	20191027000468	2	
sub122	20191028002738	2	

sub123	20191024001280	2	
sub124	20191030001526	2	
sub125	20191111002510	2	
sub126	20191126002590	4	
sub127	20191127002176	1	
sub128	20191208000592	2	
sub129	20191224000008	0	
sub130	20191228000237	0	
sub131	20160413000006	1	5
sub132	20161215001667	6	4
sub133	20200112000228	2	0
sub134	20200101000392	1	1
sub135	20201130003288	6	0
sub136	20201203002778	5	2
sub137	20201217002368	5	5
sub138	20200403000012	4	0
sub139	20200814000015	0	6
sub140	20200412000331	2	2
sub141	20200711001264	0	5
sub142	20200411000014	5	2
sub143	20200214000572	2	1
sub144	20200124000041	5	2
sub145	20200613001086	1	1
sub146	20200531000253	2	0
sub147	20201220000155	5	2
sub148	20200609001016	1	0
sub149	20200409001101	4	0
sub150	20200118000372	2	5
sub151	20201023001238	0	3
sub152	20201109000009	0	1
sub153	20201212001420	5	1
sub154	20201129000299	1	5
sub155	20200301000025	2	0

sub156	20200306000927	4	0
sub157	20201009003102	2	4
sub158	20200410001952	3	0
sub159	20200218000582	2	1
sub160	20200821002584	5	5