结论与展望

我们以脓毒症（可诱导 SCI 、 DIC ）为例，获取了 MIMIC-IV数据集，从中选择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%）。

我们分别以时间段长度分别为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。（+解析性分析结果）

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。但同时受制于时间因素，模型创新性不足，未来可进一步优化数据处理及划分等步骤，使用更多其他的模型探索其性能，进一步提高精度。