

四川大学

硕士学位论文

(工程类专业学位)

题 目 基于 MIMIC-IV 数据库的

心衰患者死亡风险预测模型

作 者 雷英鹏 完成日期 2023 年 5 月

培 养 单 位 计算机学院(软件学院)

学校指导教师 李川 副教授

企业指导教师 王涛

专业学位类别 工程硕士

专 业 领 域 人工智能与大数据

授予学位日期 2023 年 6 月

A Mortality Risk Prediction Model for Heart Failure Patients Based on MIMIC-IV Database

**A dissertation submitted to Sichuan University
in partial fulfillment of the requirements
for the degree of**

**Master of Engineering
in Artificial Intelligence and Big Data**

By

YingPeng Lei

Supervisor: Associate Prof. Chuan Li

Computer Science, Sichuan University, Chengdu, China

June, 2023

摘 要

心力衰竭（简称，心衰）是各种心血管疾病的终末阶段，已成为严重的全球公共卫生问题。目前在心衰预后中存在如下问题：（1）心衰患者病程复杂，相关因素多，难以准确评估患者预后的风险状况；（2）基于传统机器学习、医学统计等难以有效发掘心衰的预后因素；（3）心衰危险因素的研究不足。针对上述问题，本研究在 MIMIC-IV (Medical Information Mart for Intensive Care IV) v1.0 数据库中提取 6,257 名心衰患者的电子病历数据，共提出两个深度学习模型预测心衰患者死亡风险，分别基于短周期监护数据、长周期诊断数据分别判断心衰患者的短期及 1 年内死亡风险，主要贡献包括：

（1）提出基于时频感知的循环门控（temporal frequency-aware gated recurrent unit, TFA-GRU）深度学习网络，算法基于心衰患者 ICU 入院 24 小时内高维临床数据，通过临床干预及患者多源临床特征的融合模型发掘患者潜在死亡风险，研究结果表明该方法在心衰患者的短期死亡风险预测任务上能达到 0.946 AUROC (Area Under the Receiver Operating characteristic Curve) , 0.828 AUPRC (Area Under the Precision Recall Curve) ；

（2）提出基于预训练范式的 SBFHF (Slim Bert For Heart Failure) 模型结构，预训练代理任务包含改进的掩码语言模型任务及心衰患者的诊断辨别任务，结果表明该范式下的预训练代理任务能提高约 3%的心衰患者风险识别能力；

（3）本研究基于深度学习的可解释性分析，利用多种可视化角度发掘临床特征、诊断因素对于心衰患者死亡风险的影响程度，为临床知识发现提供思路。

关键词：深度学习；时间序列；心力衰竭；预测模型；死亡风险

Abstract

Heart failure is the end stage of various cardiovascular diseases and has become a serious global public health problem. At present, there are the following problems in the prognosis of heart failure: (1) the course of heart failure patients is complex, there are many related factors, and it is difficult to accurately assess the risk status of patients' prognosis; (2) It is difficult to effectively explore the prognostic factors of heart failure based on traditional machine learning and medical statistics; (3) Insufficient research on heart failure risk factors. In view of the above problems, this study extracted the electronic medical record data of 6,257 heart failure patients in the MIMIC-IV (Medical Information Mart for Intensive Care IV) v1.0 database, and proposed two deep learning models to predict the death risk of heart failure patients, respectively, based on short-cycle monitoring data and long-cycle diagnostic data, respectively, to determine the short-term and one-year death risk of heart failure patients, the main contributions include: (1) A temporal frequency-aware gated recurrent unit (TFA-GRU) deep learning network based on time-frequency perception is proposed, and the algorithm is based on the high-dimensional clinical data of heart failure patients within 24 hours of ICU admission, and explores the potential death risk of patients through clinical intervention and multi-source clinical feature fusion model, and the results show that the method can reach 0.946 in the short-term death risk prediction task of heart failure patients AUROC(Area Under the Receiver Operating characteristic Curve),0.828 AUPRC(Area Under the Precision Recall Curve); (2) A SBFHF (Slim Bert For Heart Failure) model structure based on the pre-training paradigm is proposed, and the pre-training agent task includes an improved mask language model task and a diagnostic discrimination task for heart failure patients, and the results show that the pre-training agent task under this paradigm can improve the risk recognition ability of heart failure patients by about 3%. (3) Based on the interpretability analysis of deep learning, this study uses a variety of visualization angles to explore the influence of clinical features and diagnostic factors on the risk of death in patients with heart failure, and provides ideas for clinical knowledge discovery.

Key words:deep learning;time series;heart failure;prediction model;risk of mortality

目 录

第 1 章 绪论.....	1
1.1 研究背景及意义.....	1
1.2 国内外研究现状.....	2
1.2.1 基于统计学等方法.....	2
1.2.2 基于传统机器学习方法.....	3
1.2.3 基于深度学习方法.....	4
1.3 论文组织结构.....	5
第 2 章 数据介绍及相关技术	7
2.1 数据库选型及介绍.....	7
2.2 MIMIC-IV 数据库.....	8
2.2.1 患者信息表.....	9
2.2.2 患者出入院信息表.....	10
2.2.3 诊断编码定义表.....	10
2.2.4 患者诊断表.....	11
2.2.5 ICU 停留记录表.....	11
2.2.6 ICU 事件信息定义表.....	12
2.2.7 过程事件表.....	12
2.2.8 排出事件表.....	13
2.2.9 表格事件表.....	14
2.3 深度学习相关技术.....	14
2.3.1 循环神经网络.....	14
2.3.2 Transformer.....	15
2.3.3 Deep SHAP 深度模型解释方法	15
2.3.4 词向量转换方法.....	16
2.4 统计学方法.....	16
2.4.1 学生 t 检验.....	16
2.4.2 卡方检验.....	17
2.5 模型评估方法.....	17
2.6 文中使用工具.....	18
2.7 本章小结.....	19
第 3 章 数据预处理及分析	21
3.1 数据提取.....	21
3.2 数据分析.....	22
3.2.1 临床结局分析.....	22

3.2.2 基础属性分析.....	23
3.2.3 事件表分析.....	27
3.2.4 特殊条目分析.....	27
3.3 数据清洗.....	28
3.4 本章小结.....	29
第 4 章 基于 TFA-GRU 的心衰患者死亡风险预测模型	31
4.1 方法概述及数据预处理.....	31
4.2 模型结构及细节.....	33
4.3 实验结果.....	35
4.4 解释性分析.....	38
4.5 本章小结.....	40
第 5 章 基于 Transformer 的心衰患者死亡风险预训练模型	41
5.1 方法概述及数据预处理.....	41
5.2 模型结构及细节.....	45
5.3 实验结果.....	48
5.4 解释性分析.....	53
5.5 本章小结.....	55
第 6 章 总结和展望	57
参考文献.....	59
附录 A.....	63

表格目录

表 2.1 数据集介绍	8
表 2.2 患者信息表字段介绍	9
表 2.3 出入院信息表字段介绍	10
表 2.4 诊断编码定义表字段介绍	11
表 2.5 诊断表字段介绍	11
表 2.6 ICU 停留记录表字段介绍	11
表 2.7 事件信息定义表字段介绍	12
表 2.8 过程事件表字段介绍	13
表 2.9 排出事件表字段介绍	13
表 2.10 表格事件表字段介绍	14
表 2.11 TFA-GRU 实验环境	18
表 2.12 SBFHF 实验环境	19
表 3.1 前三诊断分布	23
表 3.2 属性统计量	25
表 3.3 GCS 字段	28
表 3.4 特征映射说明	29
表 4.1 模型训练结果	36
表 5.1 预训练阶段的诊断数据统计	42
表 5.2 不同预训练模型对比	43
表 5.3 目标数据诊断数据统计	44
表 5.4 不同类型模型性能比对	49
表 5.5 eICU 诊断表字段介绍	51

插图目录

图 2.1 MIMIC-IV 数据库概览 ^[43]	9
图 2.2 DeepSHAP 的功能概览	16
图 3.1 数据提取	21
图 3.2 心衰出院与死亡时间间隔	23
图 3.3 首末次病房类型分布	24
图 4.1 不同临床结局患者的小时记录数折线图	31
图 4.2 输入数据及指示向量构造	32
图 4.3 时频编码模块	33
图 4.4 TFA-GRU 整体结构	34
图 4.5 不同参数下模型的 DCA 曲线	36
图 4.6 信息影响对比	38
图 4.7 死亡组及生存组的模型预测风险值	38
图 4.8 前 20 项死亡风险高相关因素（绝对值）	39
图 4.9 前 20 项死亡风险相关性临床特征贡献（平均）	40
图 5.1 预训练数据提取	41
图 5.2 诊断编码出现频次	44
图 5.3 预训练数据年龄分布图	45
图 5.4 SBFHF 模型结构	47
图 5.5 预训练数据与目标研究数据间的差异	50
图 5.6 不同预训练数据规模下的 SBFHF 性能	51
图 5.7 不同输入的 DCA 曲线	52
图 5.8 不同类型患者的死亡风险输出	53
图 5.9 [CLS]的注意力分布	54
图 5.10 SBFHF 的第二层的整体注意力及局部注意力分布	55

第1章 绪论

1.1 研究背景及意义

心衰十分常见，是多种心血管疾病的终末阶段^[1]。近年来，心衰的患病和发病率持续上升，已成为全球公共卫生问题之一^[2]。在2017年，全球大约有6430万的心衰患者，且未来患者数量将逐渐扩大^[3]。人口老龄化情况的加剧也将导致心衰患者越来越多^[4-5]。中国心血管健康与疾病报告2019显示，中国35岁及以上居民的患病率为1.3%，估计有心衰患者约890万^[6]。另外，心衰患者的死亡率也很高。心衰患者的住院死亡率为5%~20%，有症状的心衰患者发作五年后的死亡率高达50%^[7]。在2021年，美国心力衰竭学会领头制定的《心力衰竭通用定义和分类》^[8] (universal definition and classification of heart failure) 对心衰进行了统一的定义，规范了心衰的诊断界限^[9]。

心衰患者人数持续增长，以及心衰治疗费用的提高将不断增加医疗机构的压力以及患者的治疗负担。由于心衰具有发病率高、病情严重、治疗成本高等特点。心衰患者的预后分析研究对降低患者的医疗成本、改善临床结局非常重要。虽然近年来对心衰研究不断深入，但心衰患者的死亡风险预测研究还需要进一步加强。有效识别高风险心衰患者能有助于患者和临床医生对治疗和治疗的策略做出明智的决定。同时对低风险心衰患者可降低心理压力，避免过度医疗^[10]。

根据《心力衰竭通用定义和分类》定义，心衰是一种临床综合征，其症状和/或体征由心脏结构和/或功能异常引起，并由利钠肽水平升高和/或肺部或全身充血的客观证据所证实。然而将利钠肽或N末端B型钠尿肽前体等生物标志物用于心衰严重程度和预后的评估^[11]，往往受到许多非心源性病变的影响，预测的准确度仍然有待提高，目前亟需研发方便快捷，非侵入式且具有更高灵敏度准确性的心衰患者死亡风险辨别方法。

心衰患者的死亡风险研究目前主要使用四种模态的数据源，分别是影像学报告或出院报告一类的文本型内容、患者的心肺断层扫描、超声心动图、胸片等医学影像信息、患者住院或门诊电子病历信息以及患者的心电图等。四种数据源所包含的信息量不同，文本类报告包含的信息较为精简且具总结性质，信息直观易理解，但容易受到医生先验知识的影响。心肺断层扫描、超声心动图等表达的信息是相对客观的，但影像资料的信息较为局部化，且需依靠影像学专家解读。心电图所含信息及特点与影像学资料类似。电子病历蕴含的信息最为丰富且复杂。本文主要基于电子病历进行心衰患者的死亡风险研究。

1.2 国内外研究现状

1.2.1 基于统计学等方法

心衰是一种临床综合征，与心血管疾病发病具有直接关系。因此，在分析心衰死亡风险时可采用心血管疾病类评分模型。心血管风险评分模型中较常用的为 QRISK^[12]，Framingham^[13]，ASSIGN (ASSessing cardiovascular risk using SIGN guidelines)^[14]等。临床评分模型大部分基于 cox 回归算法，该算法是常用统计学半参数多因素生存分析方法，不要求研究数据的正态性，能发掘与死亡风险的高相关因素。

QRISK 评分模型最早发布于 2007 年，是英国的心血管疾病风险评分方法，并且每年都会以最新的 QRSearch 数据库记录进行校准。QRISK 评分是基于 cox 回归算法，推导出 10 年内患心血管疾病的风险大小。纳入因素包括糖尿病史、慢性肾病疾病、关节炎、冠状动脉心脏病家族史、皮质类固醇、偏头痛、年龄等因素。

Framingham 评分起源于美国的弗雷明汉心脏研究，其评分基于长期随访大规模人群的方式，以寻求患病人群的共同致病因素以及疾病特征。主要研究因素包括年龄、吸烟状态、收缩压、总胆固醇、糖尿病、高密度脂蛋白胆固醇来预测成人未来 10 年内发生冠心病的可能性。

ASSIGN 评分量表是基于 cox 回归算法得到的，数据来自苏格兰的心脏健康扩展队列。其与 Framingham 有类似之处，同样纳入家族病史、吸烟状况、收缩压、总胆固醇和高密度脂蛋白胆固醇，并在实验中对比了 Framingham 评分的验证效果，结果证明 ASSIGN 评分具有一定的性能提升。与 Framingham 量表相比，首次引入了社会剥夺因素 (social deprivation)，揭示了社会因素与心衰的存在一定关系。

西雅图心衰患者模型 (seattle heart failure model, SHFM) 基于 cox 回归模型算法，利用临床症状、治疗方法、实验室检验项目共 24 项评估心衰患者 1~3 年内的死亡风险。在 1125 个心衰患者上构建模型，并从多个外部数据集验证了模型的准确性和泛化性^[15]。

MUSIC (MUerte Subita en Insuficiencia Cardiaca) 评分模型^[16]在 992 名慢性心衰患者的建模队列上基于 cox 回归模型评估心衰患者的死亡风险。在仅纳入了 10 项非侵入式测量的临床指标上得到了不错的预测水平，包括心室肥大、射血分数、房颤、既往动脉粥样硬化等指标。但是该模型开发时缺乏验证队列。

MEESSI-AHF (Multiple Estimation of risk based on the Emergency department Spanish Score In patients with Acute Heart Failure)^[17]是 2017 年西班牙急诊部门针对急性心衰患者评估 30 天内死亡风险的模型，其算法是基于前向逐步逻辑回归的。该评分纳入了 13 项临床指标，包括日常活动能力指数、收缩压、年龄、N 末端 B 型钠尿肽前体、钾离子、纽约心脏病协会评级 (new york heart association, NYHA)、肌酐、肌酐蛋白、急性冠状动脉综合症、血氧饱和度、呼吸率、心室肥厚、尿量因素。建模队列来自 4,867 名急

性心衰患者数据并在 3,229 名患者上进行了外源验证, 准确度能达到近 0.83 AUROC (Area Under the Receiver Operating Characteristic curve)。

虽然心衰患者的风险评估模型较多, 但仍然有许多评分方法未在临床中广泛使用, 这主要是由于量表的预测结果与临床期望获取信息是不同的^[18]。如 MEESSEI-AHF 能估计患者的死亡率, 但却不能基于某种明确死亡率界限下, 指导急诊科医生是否需要将患者进行入院或提前出院。另外, 常规医学统计量表的预测水平还有很大提升空间, 建模队列及验证队列的规模和质量也应进一步扩大。上述两点是目前医学评估模型较为常见的不足。

1.2.2 基于传统机器学习方法

传统机器学习方法建立在医学统计模型的基础上, 引入了更多样的基础算法类型。目前针对心衰患者死亡预测任务中, 代表的模型有树结构模型、支持向量机 (support vector machine, SVM) 等。

极致梯度提升算法 (extreme gradient boosting, XGBoost) 是基于集成学习框架的树结构模型。其通过不断生成新树拟合目前预测残差以最小化预测误差, 相比梯度提升算法, XGBoost 引入二阶泰勒公式提高拟合精度, 目标函数加入正则项控制算法复杂度等, 是心衰患者死亡风险预测任务中使用最多且效果较好的。Lu 等人^[19]使用 XGBoost 算法预测左心室射血分数以评判心衰患者的死亡风险, 为处理患者时间维度上可能发生的病情变化, 将 45 天内的数据划分为代表不同时间段的样本集, 虽然增加了计算量但减小了面对长周期数据时使用简单统计量作为算法输入的不准确性。Li 等人^[20]面向心衰患者死亡风险预测目标, 基于 XGBoost 算法将患者入院后 24 h 内 ICU 的临床记录通过统计及机器学习方法预筛选, 最终利用 24 项临床记录的统计量作为模型输入, 在 2,798 名心衰患者数据上得到了较高的预测准确度, 类似的研究方法如 Luo 等人^[21], 但其通过外源数据集进行了性能验证。Zaman 等人^[22]在仅 299 名心衰患者的数据上利用 12 种临床特征建立算法, 其中 Xgboost 预测性能高于决策树、随机森林、聚类算法的结果。Takahama 等人^[23]在 987 名心衰患者数据上通过 LightGBM (Light Gradient Boosting Machine) 算法建模, 利用了尿素氮、收缩压、红细胞比容等 15 项临床特征预测一年内的患者死亡事件或安装左心室辅助装置事件的发生, 算法达到了最高 0.87 AUROC 的预测准确度, 并在仅使用肌酐蛋白、收缩压、身体质量指数、脑钠肽水平、红细胞比容共 5 项临床特征的前提下能达到 0.86 AUROC 的预测水平。LightGBM 算法是基于树结构以及集成思想的新兴算法, 在性能上更优于 XGBoost 算法。

SVM 是寻求最大化分类超平面距离的算法。姚丽娟等人^[24]基于 6,260 条患者住院记录的 1,302 条预定义临床特征数据, 通过 Relief 方法筛选高维临床特征, 并结合带偏惩罚的支持向量机 (biased penalties-SVM, bp-SVM) 模型预测 30 天内心衰患者的死亡概

率, 该模型缓和了数据集标签不平衡的问题, 虽然在构建特征时引入了较多人工判断, 但该算法是基于国内临床数据库研究的一次尝试。

传统机器学习方法也能与其他领域的方法进行结合, 预测心衰患者的死亡风险。Tunc 等人^[25]将基于关联规则的分类模型应用到心衰患者的死亡风险预测中, 将关联规则的频繁项集作为算法输出的解释性说明, 关联规则的置信度作为患者的死亡风险概率, 是数据挖掘方法与机器学习相结合的一次有效实践。Park 等人^[26]使用了 cox 回归模型与集成思想结合的 cox-boost 算法来预测急性心衰患者 3 年内的死亡风险, 基于 8 项超声特征及 19 项临床特征下达到 0.76 AUROC 的验证精度, 并通过可视化模型风险输出与实际患者的生存时间等方式证明了模型的合理性。

传统机器学习在较少样本, 特征维度低的情况下常拥有较好的模型性能。同时, 传统机器学习方法的解释性较好, 得益于算法设计常基于较成熟的理论支持, 如树结构划分来自信息论中的信息熵、聚类方法基于易理解的距离度量准则、SVM 基于最大化分类超平面距离等。但在数据规模、数据维度增大时传统机器学习方法将很难支撑。

1.2.3 基于深度学习方法

近年来深度学习在临床领域研究越来越多, 通过深度学习方法预测心衰患者的死亡风险存在多项优势。基于深度学习方法不需要过多的临床相关知识、更简化的特征工程以及更优的预测性能。基于深度学习开发的算法更有利于知识的发现, 这也是常规的统计方法和传统机器学习方法依赖先验知识、难以处理高维数据、难以发掘多维数据间关联所导致的不足。

循环神经网络 (recurrent neural network, RNN) 是较为普遍的电子病历数据处理方法^[27]。Che 等人^[28]认为不同采样频率的临床特征差异会在数据聚合时产生大量的数据缺失问题, 提出基于缺失模式以及衰退更新的循环门控 (gated recurrent unit, GRU) 网络更新策略。通过算法将临床特征的缺失模式进行追踪, 有效的将缺失信息模式和模型训练融合, 从另一种角度处理了聚合数据时引发的缺失问题, 也避免了盲目填充数据时的可能风险, 并在患者死亡风险任务预测中达到了不错的水平。刘卢琛等人^[29]对电子病历数据的不同事件类型进行独立编码, 在长短期记忆 (long short term memory, LSTM) 网络中引入事件门和时间门, 通过额外的两种门控感知不同临床事件的采样频率以控制产生网络信息流更新, 每个时间步输入仅纳入单一临床事件, 不但回避了数据对齐的问题, 也增强了算法在时间维度处理上的细粒度。

卷积神经网络 (convolution neural network, CNN) 对电子病历数据同样适用。Nguyen 等人^[30]将 CNN 的归纳偏置引入电子病历特征提取, 将患者电子病历中包含的诊断及过程手术的对应该编码所构成的序列视作模型输入, 以卷积核窗口内包含的数个编码视作临床主题 (clinical motifs) 概念, 并通过卷积核梯度发现了一系列与患者临床结局高相关

的临床主题。Li 等人^[31]用不同大小的一维卷积核与患者实验室检查、手术、诊断、常规生理体征等信息以及标识填充的指示向量所构成的一维表示向量进行卷积运算，将不同卷积核提取特征通过多头注意力机制交互不同局部窗口信息，在 10,311 名心衰患者的 30 天、180 天、360 天、360 天外的四种死亡预测任务上达到了不错的水平。

Transformer^[32]是基于自注意力机制的编解码深度模型，近年来基于 Transformer 的电子病历研究越来越多。Rasmy 等人^[33]首先归纳电子病历与自然语言的异同，将成功应用于自然语言的 Transformer 应用到电子病历数据的特征提取，提出基于 Transformer 编码器的 Med-BERT 深度模型，该模型在 28,490,650 名患者数据上通过掩码语言模型 (masked language model, MLM) 任务以及自定义的住院延长 (prolonged length of stay, prolonged LOS) 预测任务预训练模型。该方法证明了基于 Transformer 结构的预训练方式能有效提高电子病历数据的表示学习能力。

1.3 论文组织结构

本文结合深度学习方法，利用重症监护室 (intensive care unit, ICU) 中患者电子病历数据，识别高死亡风险的心衰患者，并发掘与死亡风险高相关的临床特征。从知识发现及辅助辨别风险两个角度开发算法。分别提出基于 ICU 内心衰患者 24 小时监护数据的短期时序风险预测模型，以及基于历史诊断及患者年龄判断 365 天内患者死亡风险的预测模型。两种模型都能辅助临床医生识别高风险心衰患者，更早的介入针对治疗，尽可能改善患者临床结局。

第一章：绪论。介绍心衰的临床定义及背景知识，概括心衰患者的预后死亡风险预测意义以及目前国内外与心衰患者死亡风险预测相关的统计学、传统机器学习、深度学习方法，对比方法间的优缺点，并对论文整体结构简短说明。

第二章：数据介绍及相关技术。阐述数据源选择依据、使用的数据表结构、研究方法的相关技术说明，具体包括统计方法、深度学习方法、机器学习可解释性方法。

第三章：数据预处理及分析。介绍数据提取、数据清洗、数据分析等过程。数据清洗包括异常值、缺失值处理方法、部分条目合并处理方法。数据分析阶段包括患者死亡结局分析，数据特征分析。

第四章：基于 TFA-GRU 的心衰患者死亡风险预测模型。介绍 TFA-GRU (Temporal Frequency-Aware Gated Recurrent Unit) 算法模型。模型结合电子病历中的干预记录频次、入院时间间隔、患者诊断，生理体征及临床评分量表等多源异构信息共同预测患者死亡风险，并基于解释模型预测结果，发掘临床特征与死亡风险的相关性及重要特征，为 ICU 内心衰患者提供了短时间的死亡风险评估方法。

第五章：基于 Transformer 的心衰患者死亡风险预训练模型。介绍基于 Transformer 预训练范式的心衰患者表示学习 SBFHF (Slim BERT For Heart Failure) 模型，并提出针

对心衰患者的新型预训练代理任务。结果证明提高心衰患者电子病历数据的表示学习能力。算法结合了历史诊断及患者年龄，以回顾性方式评估患者一年内的死亡风险，并通过可视化注意力权重等解释性分析，说明强相关的诊断信息。

第六章：总结与展望。归纳本文主要研究内容及方法，提出目前方法存在的局限性，为未来的研究提出一些方法思路。

第2章 数据介绍及相关技术

2.1 数据库选型及介绍

在针对心衰研究前应选择可靠且开源数据。表 2.1 列举了备选数据集，且均包含心衰病人的相关生理数据，其中心衰患者数或由于研究数据筛选差异存在不同或数据库更新，该数值仅供参考，诊断类型描述中包含国际疾病诊断分类 (international classification of diseases, ICD) 的相应版本或其他类型编码。MIMIC-III (Medical Information Mart for Intensive Care III) ^[34]数据集来源于贝斯以色列女执事医疗中心 (beth israel deaconess medical center, BIDMC) 中所收录病人数据信息，是 2015 年发布后广泛使用的医疗数据集，其包含电子病历中提取的大量临床数据，以及床旁监护设备所采集到的波形数据、体征参数、事件记录等。库中共涵盖大约 6 万条病人的 ICU 住院记录，覆盖记录时间跨度从 2001~2012 年。eICU^[35]数据发布于 2018 年，共包含超过 20 万病人的住院记录，且病人数据来自美国的 20 家医疗中心，时间跨度在 2014~2015 年之间。CPRD (Clinical Practice Research Datalink) ^[36]是英国的临床数据集，数据来自英国的全科医生网络中收集的匿名病人数据，目前包含近 6 千万的病人记录，其中涵盖 1.8 千万目前注册的病人。使用 read^[37]编码对疾病诊断、症状、生活习惯进行描述，而 read 编码并非国际通用形式。BioBank^[38]数据集是英国的生物医学数据库，包含了近 50 万病患的基因信息和健康信息，时间为 2006~2010，患者年龄 40~69 岁之间，且男女性数量基本平衡。该数据集目前是世界上最大的人类遗传队列生物样本库，并结合英国的电子医疗数据。其中包含人口统计、基因、生物标志物、随访、健康结局等相关数据。根据官网 2023 年 3 月数据显示，心衰患者的登记报告数为 21,011 名^[39]。Ahmad 等人^[40]提出基于心衰病人的小型数据集 (heart failure clinical records data set, HFCRDS)，其患者年龄都在 40 岁及以上并存在左心室收缩功能不全的问题，在 NYHA 评级为 3 和 4 级，共包含 299 名心衰患者，可从 UCI (University of California Irvine) 数据集网站获取，其中纳入了贫血状况、肌酐、血小板数、钠离子、射血分数等 13 项特征，并未有任何数据缺失。MIMIC-IV^[41]是国外开源的临床数据库。数据主要来自 BIDMC，分为全院级别的电子病历和重症监护病房患者的数据，全院级别数据涵盖 38 万病人，ICU 数据包含 5 万 3 千多名患者。MIMIC-IV 是 MIMIC-III 的后续版本，加入更新的患者数据。数据包含诊断、药物、实验室检验、人口统计、护理人员记录等多源临床信息。包含数据类别广，涵盖信息全的特点使得 MIMIC 系列数据能有效推动多方向的临床科研、提升患者治疗质量、改进算法性能等，并增加了胸部摄片数据和急诊科的数据。ADHERE (The Acute Decompensated Heart Failure National Registry) ^[42]数据集是 2003 年美国为解决急性失代偿性心衰的住院问题，

建立急性失代偿性心力衰竭患者数据库，由主或次出院诊断包含心衰为纳入依据，收集了 263 家医院患者的多项医疗历史记录。

综合上述数据集特点及来源，选择 MIMIC-IV 数据库作为研究对象有以下原因；

- 1) MIMIC-III 的数据研究较为成熟，衍生的 MIMIC-IV 具有相似结构，可参考过去研究，
- 2) MIMIC-IV 的数据集记录数相对合适，有利于机器学习和深度学习模型的学习，
- 3) 发布于 2020 年，是较新的数据，4) 数据获取相对便利，只需完成线上考试认证。

表 2.1 数据集介绍

Table 2.1 Introduction to datasets

数据集名	发布年	总纳入患者	诊断类型	心衰患者数
eICU	2018	200,000+	ICD-9/10	17,029
MIMIC-IIIv1.4	2016	60,000+	ICD-9	10,436
MIMIC-IVv1.0	2021	382,278	ICD-9/10	25,467
BioBank	2015	500,000+	ICD-10	21,011
CPRD	2015	60,000,000	read	-
HFCRDS	2017	299	-	299
ADHERE	2003	65180	-	65,180

2.2 MIMIC-IV 数据库

MIMIC-IV 数据库最早发布于 2020 年，本文主要基于 2021 年的 v1.0 进行数据提取。MIMIC-IV 中由 `mimic_hosp`、`mimic_icu`、`mimic_core`（分别简称 `hosp`、`icu`、`core`）三个模块构成，共 27 张数据表。“`icu`”是“`hosp`”中提取在 ICU 停留过的患者记录，囊括更为详细的院内体征监护、检测指标等信息，目前数据库整体结构如图 2.1^[43]。其中 `Metavision` 为 ICU 的信息系统数据，`external` 为外部搜集数据，`note` 模块为 v2.0 版本后中加入的文本报告模块，包括放射学报告以及出院报告等，在 v1.0 中并未引入。`core` 模块包含 `patients`（患者信息）、`admissions`（出入院信息）、`transfers`（病房转移信息）三张表，该类表唯一确定了患者、患者住院、院内转病房的唯一标示号以及相关的时间等信息，可用于查询目标患者关联数据。“`hosp`”是住院病人的信息模块，包括了诊断（`diagnoses_icd`）、药房数据（`pharmacy`）、药物管理（`emar`、`emar_detail`）、实验室检验（`labevents`）、处方数据（`prescriptions`）、手术信息（`procedure_icd`）、微生物（`microbiologyevent`）、电子记录信息（`poe`、`poe_detail`）、疾病诊断相关分组（`drgcodes`）、服务类型（`services`）、医疗保健通用程序编码（`hcupsevents`）、各项编码定义查询表（`d_hcpcs`、`d_icd_diagnoses`、`d_icd_procedures`、`d_labitems`）。“`icu`”是记录 ICU 病房内产生的信息，包括患者用药事件记录（`inputevents`）、患者排出事件记录（`outputevents`）、过程事件记录（`procedureevents`）、

表格事件记录 (chartevents)、时间事件记录 (datetimeevents) 以及 ICU 停留记录 (icustays)。本文主要以 ICU 作为主要数据来源, 探究 ICU 内的心衰患者死亡风险, 后文共使用 9 张表数据, 下文将详细介绍使用的数据表字段特点及相关含义用途。

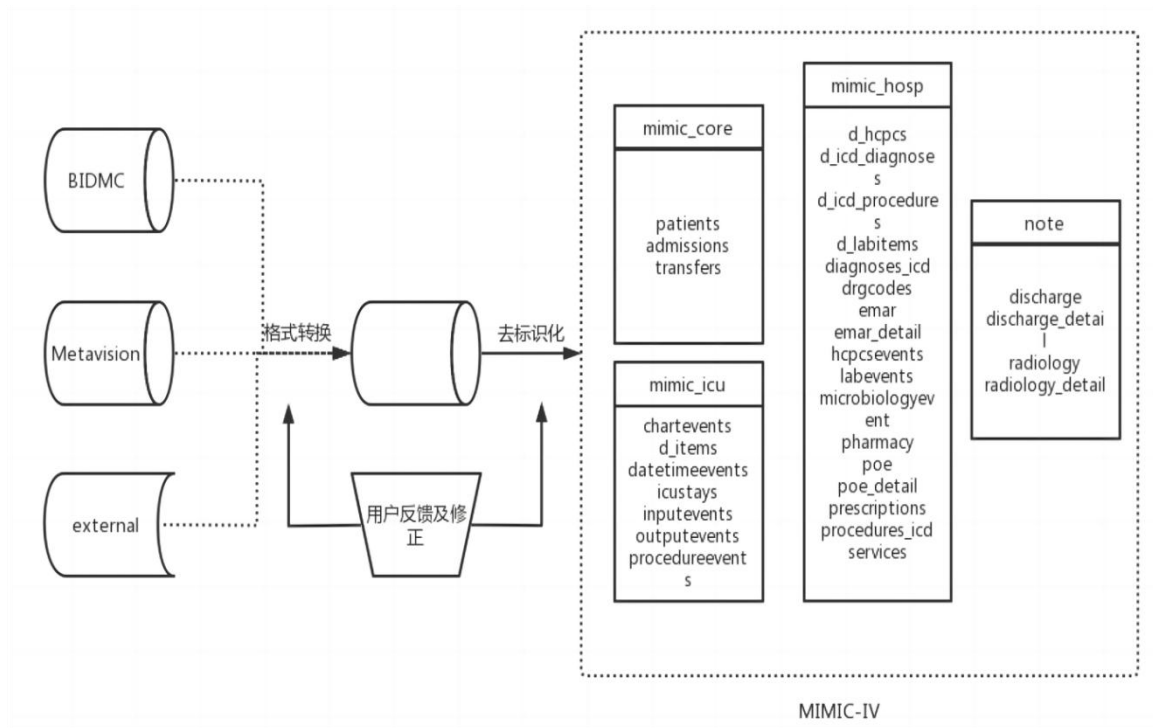


图 2.1 MIMIC-IV 数据库概览^[43]

Figure 2.1 MIMIC-IV database overview^[43]

2.2.1 患者信息表

患者信息表主要包含患者的人口统计信息, 中英文字段对照见表 2.2。其中患者唯一标识号字段独立标识一个患者, 多次入院不会变化。患者年龄若大于 89 岁统一以 91 岁记录。去标识化入院年经过处理, 可由实际入院年份区间推断大致入院时间范围, 需要注意的是, 实际入院年区间在数据中仅包含 4 种, 分别是“2014 - 2016”、“2017 - 2019”、“2008 - 2010”、“2011 - 2013”, 但相应的去标识化入院年与入院年区间并非是一一对应, 即“2014 - 2016”包含三年, 但去标识化入院年对应了 95 个数字。不同实际入院年区间的相同去标识化入院年并没有表示实际意义, 不能具体推测出时间具体是哪一年。死亡日期仅记录了患者最近次入院内的相关记录, 无院外随访记录。

表 2.2 患者信息表字段介绍

Table 2.2 Description of patients fields

名称	含义	名称	含义
subject_id	患者标识号	anchor_age	保险类型

表 2.2 续

Table 2.2 Continued

名称	含义	名称	含义
gender	住院标识号	anchor_year	语言
dod	入院时间	anchor_year_group	婚姻状况

2.2.2 患者出入院信息表

该表包含所有患者的历史出入院信息，中英文字段对照见表 2.3。住院标识号唯一标识患者的住院情况，一名患者可有多个住院标识号即多次入院记录，同时包含部分人口统计学信息。由于可能存在数据时间记录错误问题，入出院时间可作为院内其他记录时间正确性检验的重要标准。另外，是否院内死亡标签仅标注当次住院标示号对应的时间段内是否发生死亡，与患者其他住院是无关的。入院类型包括急诊入院、观察入院、直接观察等 8 种类型。入院来源由门诊手术转院、诊所转院等 10 种情况。

表 2.3 出入院信息表字段介绍

Table 2.3 Description of admissions fields

名称	含义	名称	含义
subject_id	患者标识号	insurance	保险类型
hadm_id	住院标识号	language	语言
admittime	入院时间	marital_status	婚姻状况
dischtime	出院时间	ethnicity	种族
deathtime	死亡时间	edregtime	急诊入院时间
admisstion_type	入院类型	edouttime	急诊出院时间
admisstion_location	入院来源	hospital_expire_flag	是否院内死亡
discharge_location	出院去向		

2.2.3 诊断编码定义表

表 (mimic_hosp.d_icd_diagnoses) 中描述数据库中所有的 ICD 编码，中英文字段对照见表 2.4。在库中涵盖了 ICD-9 和 ICD-10 两个版本，该表的疾病诊断编码与通用的 ICD 编码是一致的。在缺乏 ICD 编码相关知识的前提下，可利用该表的诊断编码说明筛选包含目标疾病相关描述的 ICD 编码，注意在使用关键字模糊匹配时需考虑介词伴随 (with) 及未伴随 (without) 的否定句式存在。

表 2.4 诊断编码定义表字段介绍

Table 2.4 Description of d_icd_diagnoses fields

名称	含义	名称	含义
icd_code	疾病诊断编码	icd_version	编码对应版本
long_title	诊断编码说明	-	-

2.2.4 患者诊断表

该表是患者入院后产生的一系列诊断编码的汇总信息，详细中英文字段对照见表 2.5。主次诊断顺序可判断症状的主次级别，以划分患者的主次病症。该表配合诊断编码定义表可筛选目标患者住院阶段的诊断信息。

表 2.5 诊断表字段介绍

Table 2.5 Description of diagnoses_icd fields

名称	含义	名称	含义
subject_id	患者标识号	hadm_id	住院标识号
seq_num	主次诊断顺序	icd_code	疾病诊断编码
icd_version	编码对应版本	-	-

2.2.5 ICU 停留记录表

表包含所有 ICU 患者期间的转入转出时间、监护室类型、停留时间等，中英文字段对照见表 2.6。住院时间 (length of stay, LOS) 是患者在 ICU 总住院时间，以天为单位。ICU 病房类型包含 9 种，分别为冠心病监护室 (coronary care unit, CCU)、外科重症监护室 (surgical intensive care unit, SICU)、神经降压病房 (neuro stepdown)、神经过渡病房 (neuro intermediate)、神经外科重症监护室 (neuro surgical intensive care unit, Neuro SICU)、心血管重症监护室 (cardiac vascular intensive care unit, CVICU)、创伤外科重症监护室 (trauma surgical intensive care unit, TSICU)、内科重症监护室 (medical intensive care unit, MICU)、内科/外科重症监护室 (MICU/SICU)，对应记录在首末次 ICU 病房类型中。通过该表的 ICU 标识号可唯一跟踪患者 ICU 内所有记录信息。

表 2.6 ICU 停留记录表字段介绍

Table 2.6 Description of icustays fields

名称	含义	名称	含义
subject_id	患者标识号	last_careunit	末次 ICU 病房类型

表 2.6 续

Table 2.6 Continued

名称	含义	名称	含义
hadm_id	住院标识号	intime	转入时间
stay_id	ICU 标识号	outtime	转出时间
first_careunit	首次 ICU 病房类型	LOS	住院时间

2.2.6 ICU 事件信息定义表

表 (mimic_icu.d_items) 定义“icu”全部测试项的自定义检验项编号，以及具体检验的说明，中英文字段对照见表 2.7。“icu”中表组织是星型结构，该表作为中心的定义表，与其余事件表紧密结合，表中的检验项编号十分重要。如在查询体温值时需先通过该表预先找到检验项的事件编码，可通过事件描述字段过滤。然后通过关联表确定体温检验项纳入的表空间，最后在纳入表空间中筛选检验项编号值的数据项即可。该表中的正常低值和正常高值缺失较为严重，可不考虑两列的数值状况。事件类型包括常规生理体征、临床重大事件、饮食、呼吸系统、手术等 79 种。

表 2.7 事件信息定义表字段介绍

Table 2.7 Description of d_items fields

名称	含义	名称	含义
itemid	检验项编号	category	事件类型
label	事件描述	unitename	事件值单位
abbreviation	事件缩写	param_type	事件值类型
linksto	关联表	lownormalvalue	正常低值
highnormalvalue	正常高值	-	-

2.2.7 过程事件表

该表主要包含患者的手术信息、参与的治疗检测手段记录、重大事件记录等，中英文字段对照见表 2.8。其中检测手段包括 CT、胸片等是否有检查过，实际报告和影像学资料需在 mimic-cxr 数据模块中获取。该模块独立于 MIMIC-IV 主模块，需要额外获取。重大事件描述患者是否经历了如胸痛、房颤、开胸手术、摔倒等一系列事件。另外该表还描述了过程的执行状况，分为结束，完成，暂停三种。由于手术、插管、透析等操作是持续的，因此该表记录时需额外记录开始和结束时间。总剂量及相应单位、过程取消原因、备注日期列在数据库中并未存在记录。而值列针对重大事件为 1 时表示发生，针

对插管或持续治疗等状况与值单位共同表示持续时间，如值为 30，值单位为 min 则表示该事件持续 30 分钟。条目类型描述包含 Task 以及 ContinuousProcess 两种类型，表示该项过程是持续过程还是行为出现的描述，当条目类型描述为 Task 时 value 只能为 1。条目标示号及关联条目号为标注完成同一事件，但由于暂停等原因分为多次执行时，继续执行的新条目中关联条目号为上次暂停的条目标示号，当本条目没有关联条目时取自己的条目号作为关联条目号。

表 2.8 过程事件表字段介绍

Table 2.8 Description of procedureevents fields

名称	含义	名称	含义
subject_id	患者标识号	orderid	条目标示号
hadm_id	住院标识号	linkorderid	关联条目号
stay_id	ICU 标识号	ordercategoryname	条目类型名
starttime	开始时间	secondaryordercategoryname	二级条目类型名
endtime	结束时间	ordercategorydescription	条目类型描述
storetime	登记时间	totalamount	总剂量
itemid	检验项编号	totalamountuom	总剂量单位
value	值	originalamount	原始剂量
valueuom	值单位	originalrate	原始速率
location	部位	cancelreason	取消原因
locationcategory	部位类型	statusdescription	状态描述
patientweight	体重	comments_date	备注日期
Isopenbag	输液袋是否打开	continueinnextdept	出科后是否续药

2.2.8 排出事件表

表记录了患者不同引流管排出的液体记录，中英文字段对照见表 2.9。

表 2.9 排出事件表字段介绍

Table 2.9 Description of outputevents fields

名称	含义	名称	含义
subject_id	患者标识号	storetime	登记时间
hadm_id	住院标识号	itemid	检验项编号
stay_id	ICU 标识号	value	记录值（字符型）
charttime	记录时间	valueuom	单位

表中记录了收集液体量与导管的名称或类型，需要结合临床知识确定引流管位置。

2.2.9 表格事件表

该表记录患者床旁监护设备数据、治疗设备配置、护理计划、治疗手段、日常事件、实验室检验项、评估量表等患者多侧信息，中英文字段对照见表 2.10。医生标注字段数值较少且表中说明不足，可忽略该列。表格事件是“icu”中数据纳入类型最为广泛的，记录了生理体征等重要患者特征，因此在后续研究中将该表作为主要数据来源。该表存在共 2,226 项检验项。在后续使用时需要具体分析表中数据构造并结合临床经验，防止错误使用检验项条目。

表 2.10 表格事件表字段介绍

Table 2.10 Description of chartevents fields

名称	含义	名称	含义
subject_id	患者标识号	storetime	登记时间
hadm_id	住院标识号	itemid	检验项编码
stay_id	ICU 标识号	value	记录值（字符型）
charttime	测量时间	valueuom	数值单位
warning	医生标注	valuenum	数值

2.3 深度学习相关技术

2.3.1 循环神经网络

RNN 与一般神经元不同处在于结构中存在循环连接。循环连接以及神经元的状态计算逻辑使得网络具有记忆能力。然而在 RNN 模型中，当序列长度超过某一范围时，循环连接的存在将导致神经元的梯度计算随着长度线性倍增或倍减，网络可能忘记较早历史信息，这也称为梯度爆炸或梯度消失问题。为了缓解网络梯度问题，目前使用更多的是基于 RNN 的 LSTM 或 GRU 结构，二者因加入门控机制以及新的激活函数使得深度网络具备更长期的记忆性。其中 LSTM 是在 RNN 中加入三类门控单元，分别是输入门、输出门和遗忘门，三者共同控制网络输出以及存储的记忆信息流，并使用 sigmoid 激活函数抑制了梯度爆炸的问题，但仍然可能存在梯度消失。后来，又提出了 GRU 结构，仅采用重置门（reset gate）和更新门（update gate）控制信息流，相比 LSTM 减少了一定参数量，并且在实验中证明了模型表现能力并无差异。目前 RNN 一般可泛指 LSTM 或是 GRU 等相关变体结构，该类网络善于处理存在时间维度变化的数据，一般称为时序数据，电子病历可看作患者的时序数据。在心衰研究领域内，RNN 主要用于

电子病历数据、临床文本报告、心肺音频信号、心电图等的特征提取，其优点是能在不同长度的序列数据上学习，具有动态建模的能力而无需改变网络结构。最初 RNN 只能捕捉单向顺序信息，在文本报告中需要结合上下环境感知，因此研究者提出了双向循环神经网络（bidirectional recurrent neural network, Bi-RNN），以在当前时间步同时结合过去（前）和未来（后）信息，以提高全局信息感知能力。

2.3.2 Transformer

Transformer 是经典的基于注意力机制模型。为解决早期序列模型的计算并行困难，提取特征信息受限序列长度所提出。网络由解码器和编码器两个部分构成。编解码器又分为数个 Transformer 块，每个 Transformer 块由两个子模块构成，分别是多头注意力模块以及前馈神经网络模块。其中多头注意力模块根据编解码阶段不同可分为自注意力模块以及带掩码的注意力模块。多头注意力模块中，注意力基于可缩放的点积运算，而多头指通过不同线性层投影后的计算单元。传统注意力机制的计算即可理解为两个对象之间的关联匹配度，在 Transformer 块中通过三个不同的投影层将输入向量映射为查询向量（query）、配对向量（key）、值向量（value），并将查询向量和配对向量的可缩放点积运算得到的注意力作为值向量的加权乘数。带掩码的注意力模块加入了掩码机制防止如机器翻译等场景下，翻译过程中不应关注后文信息所设计的特殊处理手段。由于 Transformer 最早提出是为了解决机器翻译等问题，后续 BERT（Bidirectional Encoder Representations from Transformers）^[44]将 Transformer 应用到更高层级的预训练范式，堆叠编码器的 Transformer 块作为特征提取器，在不更改任何结构的前提下，仅通过预定义的 MLM 任务和句子预测任务（next sentence prediction, NSP），进一步提升了自然语言领域的多项下游预测任务的性能。MLM 任务通过遮罩句中部分词或随机替换等方式添加噪声，利用掩码位置的预测损失修正编码器的上下文感知能力。NSP 任务则是输入一对配对或未配对的句子，通过模型预测是否上下句匹配方式以提高模型整体语境的感知能力。MLM 及 NSP 任务大大增强了编码器的特征健壮性，也减少了数据标注的需求。BERT 是基于大规模数据的预训练范式，不但可以解决如临床领域的小样本问题，其中的预训练代理任务也进一步降低了算法的使用门槛，避免了人工标注的烦琐和有标注数据集的缺失。

2.3.3 DeepSHAP 深度模型解释方法

DeepSHAP^[45]是目前深度学习的常用解释方法，方法起源于博弈论。其假设深度模型当前样本输出与总体预测期望的偏差由该样本各个特征与结局的正负贡献积累而成，该正负贡献用 shapley 值的大小表示。DeepSHAP 无需考虑深度模型的内在结构，将模型本身视作一种解释性模型，而将模型的任意预测看做一种解释性表达方法从而解耦了

模型结构。能发掘患者预测结果与输入特征的正负关联，生成深度模型预测的解释性决策路径。同时 DeepSHAP 框架具备不错的计算性能以及更好地解释一致性，更适用于解释复杂的模型，该框架可简单理解为图 2.2，其中的基础风险由训练数据计算得到。

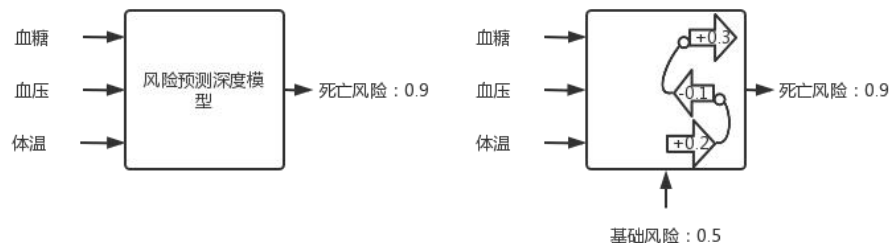


图 2.2 DeepSHAP 的功能概览

Figure 2.2 Overview of DeepSHAP's features

2.3.4 词向量转换方法

计算模型需要将字母或文字等不同情境下的基础词源转换为计算机可感知的数值向量。最早使用的词袋模型 (bag of words) 将目标数据集中的单词收集汇总，给予每个单词独一无二的数值标识号。参与计算时，句中的单词需使用词表大小的向量表示，或使用整个句子的单词统计量表示。该方法表示向量一般极度稀疏，没有实际语义表达，无法分辨各个词语之间的关联。

Word2vec^[46]方法是在词袋模型基础上，基于大规模语料库中上下文词源的关联，通过两种不同的训练方式得到目标词的表示向量，向量维度一般远小于词表大小，并且在训练后能得到表示空间中的抽象语义信息，具备通过表示向量得到近义词的语义感知能力。具体训练方式分别为连续词袋 (continuous bag-of-word, CBOW)、跳词 (skip-gram)，两种方法都预先选择一个窗口大小内的词，分别用周围词预测中间词或中间词预测周围词。

2.4 统计学方法

2.4.1 学生 t 检验

学生 t 检验 (Students's t-test) 理论上用于总体呈正态分布的样本数据。主要分为三种检验方式，独立样本检验、配对样本检验、单样本检验。独立样本 t 检验主要评价两个互相独立的样本之间的某属性差异。配对样本 t 检验的样本来自同一个体或互相配对的两个样本对。单样本 t 检验用于评判样本均值与已知总体之间的差异水平。检验方式差异源自研究对象之间的关系。t 检验的使用前提包括方差齐性和数据的正态性。在检

验样本量相对较大时由于中心极限定理可近似忽略前提条件。

统计学上，常使用假设检验思想判断命题的合理性。t 检验的原假设即两组样本数据的均值相等，在该假设前提下判断该命题的发生概率，当显著性水平或称发生概率小于某值时一般认为是不太可能发生的小概率事件，则可否定原假设。独立样本 t 检验计算公式如下：

$$t = \frac{|\tilde{x}_1 - \tilde{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (2-1)$$

式中： $\tilde{x}_1, \tilde{x}_2$ —— 组内样本均值； n_1, n_2 —— 组内样本数； s_1, s_2 —— 组内标准差

当该值越大说明两组的均值差异越大。后文中采用独立样本 t 检验方法，用于评判不同临床终点的患者组间是否在某临床特征下存在显著性差异，而该差异水平可用于特征筛选时的参考依据。显著性评判需根据样本自由度即（样本量大小-2）以及期望的显著性水平共同判定。

2.4.2 卡方检验

卡方检验针对类别型离散样本，用于统计样本实际观测值与理论值的偏离程度。卡方值越大，偏离程度越大，相应类间的差异越大。根据卡方检验结果可判断离散特征在满足某类别划分下的重要性。计算公示如下：

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (2-2)$$

式中： O_i —— 观测值大小； E_i —— 代表理论值大小

其中理论值一般由均匀分布下，类别样本数对应采样总样本数的占比计算得来。当样本存在期望频数<1 时，卡方检验失效，应改用 Fisher 精确检验替代。

卡方检验在使用时需要判别样本的自由度以及对应的显著性阈值。自由度计算根据（类别数-1）·（行数-1）。根据显著性水平可确定卡方检验的特征是否与患者结局存在高相关度。

2.5 模型评估方法

文中采用了两种曲线下面积衡量模型性能，分别是 AUROC 和 AUPRC (Area Under the Precision Recall Curve)。在 ROC (receiver operating characteristic curve) 中，根据不同阈值下模型的假正率 (false positive rate, FPR) 和真正率 (true positive rate, TPR) 绘制曲线，纵轴为 TPR，横轴为 FPR。而 PRC (precision recall curve) 是在不同判别阈值下召回率 (recall) 和精确率 (precision) 的度量，纵轴为 precision，横轴为 recall，相关计算见（公式 2.3 ~ 2.5）。

$$FPR = \frac{FP}{FP + TN} \tag{2-3}$$

$$FPR = \frac{FP}{FP + TN} \tag{2-4}$$

$$Precision = \frac{TP}{TP + FP} \tag{2-5}$$

假设当正负例均衡时，增加大量负例数量，TPR (Recall) 只与正例相关不发生变化，FPR 中分子分母都会增加，precision 中只有 FP 增加。在不断改变阈值过程绘制两种曲线时，PRC 的 precision 的变化明显高于 ROC 的 FPR，即 PRC 能在样本分布不均衡时仍然具备较好感知能力，因此同时纳入两种度量方式。

但是，两种曲线都未考虑临床结局的影响，不具备筛选更适用临床任务算法的能力。2006 年，Andrew Vickers 博士提出了决策分析方法 (decision curve analysis, DCA) [47]。DCA 的思想源自设计临床决策或风险预测模型时根据不同患者群体，分为有疾病治疗、无疾病治疗、有疾病不治疗、无疾病不治疗四种状况，假设在某判别阈值条件下划分治疗与不治疗样本组满足治疗组代价期望与避免治疗组的期望代价相同，另假设其临床代价分别是 a、b、c、d，并将 p 作为患疾病的概率，因此可以得到相关等式：

$$\begin{aligned} p \cdot a + (1 - p) \cdot b &= p \cdot c + (1 - p) \cdot d \\ p \cdot (a - c) &= (1 - p) \cdot (d - b) \\ \frac{p}{1 - p} &= \frac{d - b}{a - c} \end{aligned} \tag{2-6}$$

其中 (a-c) 是有疾病患者条件下，避免治疗相对治疗而言所带来的临床代价，(d-b) 是无疾病患者条件下，相对不治疗而言额外治疗所带来的临床代价。从公式 (2-6) 可知判断患者是否治疗的关键在于如何度量这四种情况的临床结局影响程度。基于公式 (2-6) 的分析后，Andrew Vickers 博士定义 DCA 曲线的纵轴为净收益 (Net Benefit)，横轴为判别阈值 p，公式如下：

$$NetBenefit = \frac{TruePositiveCount}{n} - \frac{FalsePositiveCount}{n} \frac{p}{1 - p} \tag{2-7}$$

式中：p——判别概率阈值；n——样本总数；TruePositiveCount——真阳例总数；FalsePositiveCount——假阳例总数

2.6 文中使用工具

文中研究使用包括硬件、软件部分。其中 TFA-GRU 网络训练使用主要环境见表 2.11。

表 2.11 TFA-GRU 实验环境

Table 2.11 Experimental environment of TFA-GRU

名称	配置信息
处理器	intel core i5 6500
硬盘	1T HDD + 128G SSD

表 2.11 续

Table 2.11 Continued

名称	配置信息
内存	16G
显卡	RTX 1050Ti (4G)
编程语言	Python3.7
数据库	Postgres13
操作系统	Windows10
编程平台	Pycharm 2021 Community

SBFHF 的训练主要环境见表 2.12。

表 2.12 SBFHF 实验环境

Table 2.12 Experimental environment of SBFHF

名称	配置信息
处理器	intel core i7 12500
硬盘	3T HDD + 512G SSD
内存	64G
显卡	RTX 3090Ti (24G)
编程语言	Python3.9
数据库	Postgres13
操作系统	Ubuntu 18.04
编程平台	Visual Studio Code

2.7 本章小结

本章首先从多个临床数据库分析比较, 介绍选择 MIMIC-IV 数据库的依据, 并详细介绍了 MIMIC-IV 数据库的整体结构及使用表的详细字段结构, 另外还结合本文研究方法说明每张表用途及注意事项, 为后续章节的数据建模基础进行了铺垫。同时, 后续研究所涉及的统计学习、深度学习、临床研究方面所涉及内容也都进行了简短的介绍。

第3章 数据预处理及分析

3.1 数据提取

MIMIC-IV中心衰患者数据需要进行预提取。本文根据心衰症状出现作为预筛选条件，在诊断编码定义表中通过含心衰症状编码说明后得到72项相关诊断编码，根据人工审核去除12项包含否定句式的非心衰编码后最终共纳入60项ICD编码（ICD-9: 25, ICD-10: 35），详细编码见附录A-1。

构建数据集时，筛选标准遵循，1) 患者年龄不小于18岁，2) 患者前三优先级诊断中包含心衰诊断编码，3) ICU停留时长不小于1天。数据提取按照如下顺序，1) 根据ICD代码定义表，仅选取与心衰相关ICD编码，2) 从患者诊断表中筛选前三诊断包含心衰编码患者，获得对应患者标识号和住院标识号，3) 根据患者标识号和住院号查询ICU病房信息，仅保留停留时间大于1天，时间最近的一条，4) 将确定的患者标识号，住院标识号，ICU标识号与各事件表进行连接查询。数据提取步骤见图3.1。

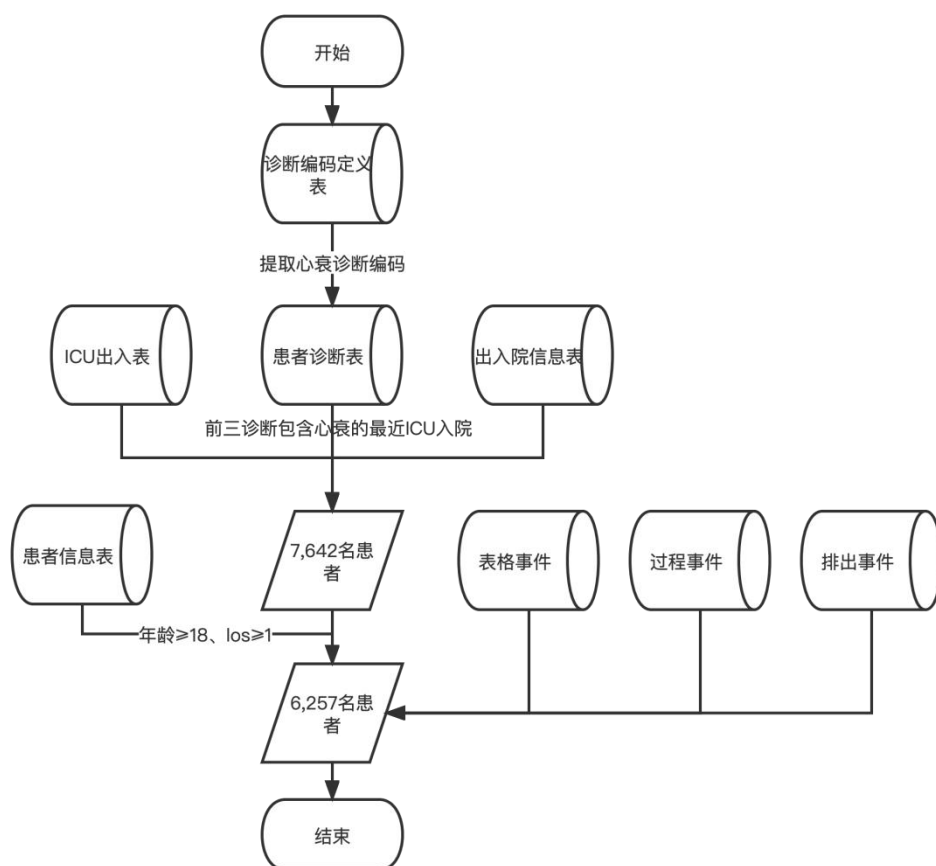


图3.1 数据提取

Figure 3.1 Data extraction

最终提取满足条件患者 6,257 名，共 26,731,762 条 icu 表数据，以及对应人口统计信息。在筛选过程中，发现记录异常数据共 828 条，该类数据与患者 ICU 停留时间段不符合，因此直接废弃。

3.2 数据分析

3.2.1 临床结局分析

首先定义住院期间的前三诊断包含心衰且死亡的患者为心衰院内死亡，否则为非心衰院内死亡，根据 3.1 节数据提取步骤可知患者临床结局可划分为心衰院内死亡与非心衰院内死亡。引入不平衡率 (imblance ratio, IR) 用于评判样本的不均衡程度，将死亡患者、非死亡患者分别视为正例 (PositiveSample)、负例 (NegativeSample)，计算如下：

$$IR = \frac{\text{NegativeSample}}{\text{PositiveSample}} \quad (3-1)$$

以心衰院内死亡作为依据时，6,257 名心衰患者中有 590 名患者满足该条件，IR 为 9.60 (5667:590)。而以诊断心衰后的非心衰院内死亡作为额外条件下，有另外 365 名患者满足，IR 为 5.55 (5302:955)，共 5302 名患者在末次出院后存活。

由于非死亡与死亡的患者数量存在不平衡状况，模型应考虑不平衡数据的处理办法。目前主要采用的方法有：通过样本加权方式给予样本不同关注度；通过降采样降低负样本比例；通过重采样增加正样本比例；通过合成数据的方法增加正样本比例。其中，样本加权对正负样例给予不同比例的训练关注度，相比基于采样的方法不容易导致过拟合或欠拟合问题，但加权比例作为超参数需多次试验，而与合成数据方法相比则操作更简单，且电子病例数据的合成合理性仍然有待探究。因此本研究采用样本加权方式，在不改变原始数据的前提下进行模型训练，保证了整体训练数据量，同时也避免了合成数据带来的可靠性问题。

另外，在患者结局衡量时，出院后的生存时间也十分重要，这对死亡风险评估具有重要意义。因此根据当前样本患者，绘制死亡患者的生存时间分布图 3.2。在 0 天范围内，包含两名特殊病人并非在心衰住院期间死亡，而是短时间内以紧急再入院的方式导致了死亡时间和目标出院间隔小于 1 天。但由于其主诊断中并未包含心衰，因此仍将其视作负例。最终，共 590 名患者在心衰住院周期内死亡，出院后 30 天内死亡患者 650 名，803 名患者在心衰出院后的一年内死亡。

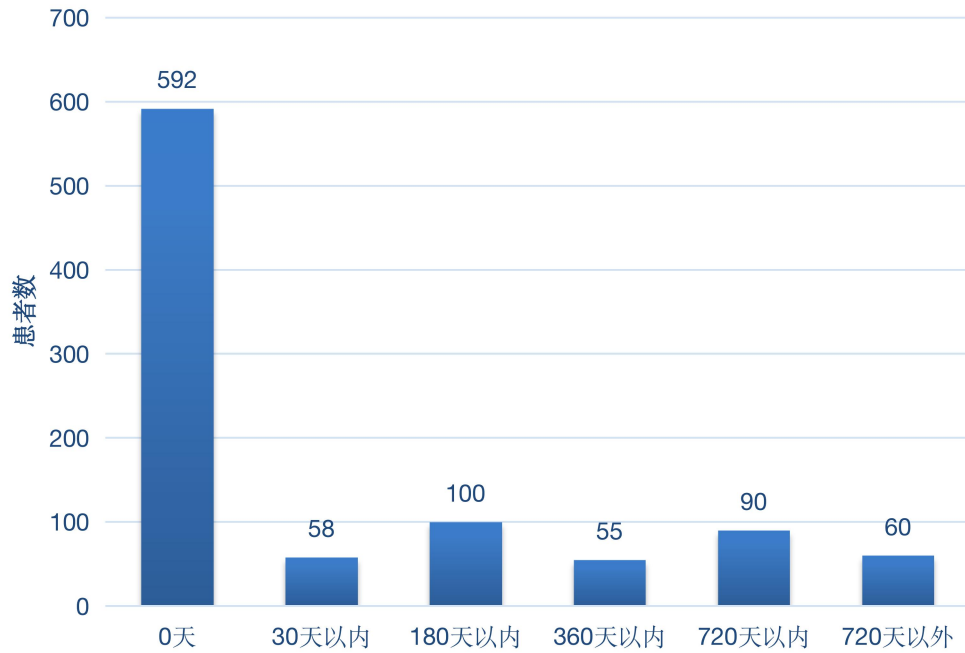


图 3.2 心衰出院与死亡时间间隔

Figure 3.2 Interval between heart failure discharge and death of date

3.2.2 基础属性分析

本节主要分析诊断、人口统计等基础信息。3.1 节提取数据时通过前三诊断包含心衰症状作为患者纳入条件，为检验目标数据集是否满足心衰患者研究标准，将该条件下提取的患者诊断数据的前三优先级分布汇总，表 3.1 描述前三诊断中出现频次较高的 ICD 类型，为便于一致性分析统一将诊断编码用 ICD-10 表示。据表 3.1 可知目标患者群体多伴随肾脏疾病、呼吸系统疾病、心脏疾病，其中心衰诊断占据了前三诊断中很大的比例，且心肌梗塞及心内膜炎的出现频次也反映了目标数据集患者多伴随心脏问题，结合前三诊断整体分布来看，可认为目标数据集符合心衰患者的研究要求。另外，伴随肾衰竭或呼吸系统疾病的患者也占一定比例，因此在对目标患者的死亡风险进行评估时，应额外注意肾脏功能、呼吸系统相关指标对于患者死亡风险的影响。

表 3.1 前三诊断分布

Table 3.1 Distribution of the top three diagnostics

第一诊断 (ICD-10) 次	第二诊断 (ICD-10) 次	第三诊断 (ICD-10) 次
心衰 (I50) 883	心衰 (I50) 3114	心衰 (I50) 1965
急性心肌梗塞 (I21) 811	肾衰竭 (N17) 329	肾衰竭 (N17) 555
心内膜炎 (I38) 378	其他呼吸道疾病 (J98) 308	呼吸衰竭 (J96) 278

数据集中，还包括了人口统计信息及院内基础信息，这些信息包括性别、年龄、婚姻状况、人种、语言类型、住院相关信息、保险信息等。其中男女性别比值为 1.23

(3453:2804)，平均年龄 71 ± 13 岁。婚姻状况包括已婚 (2,876)、未婚 (1,331)、寡妇/鳏夫 (1,298)、离异 (312)、未知状况 (440)。种族包含白种人 (4,508)、非裔美国人 (627)、拉丁人 (180)、亚洲人 (142)、印第安人 (15)、未知 (468)、其他 (256)、无法获取 (60)，将无法获取、未知、其他均合并为其他。89% (5569/6257) 患者说英语，其余未知。

院内基础信息中包含入院类型、入院来源、患者的首次 ICU 监护室类型、末次 ICU 监护室类型、医疗保险状况。其中医疗保险分为三种类型，分别是 "Other"、"Medicare"、"Medicaid"，对应人数 2,289、3,727、241。由于首末次监护室类型相对复杂，绘制柱状图 3.3 描述数据集整体分布情况。临床上，心衰作为主症状时一般会在 CCU、CVICU 停留。但从数据分布看，有一部分患者可能是住院后伴随心衰，但最初及最终的主要症状并非心衰。

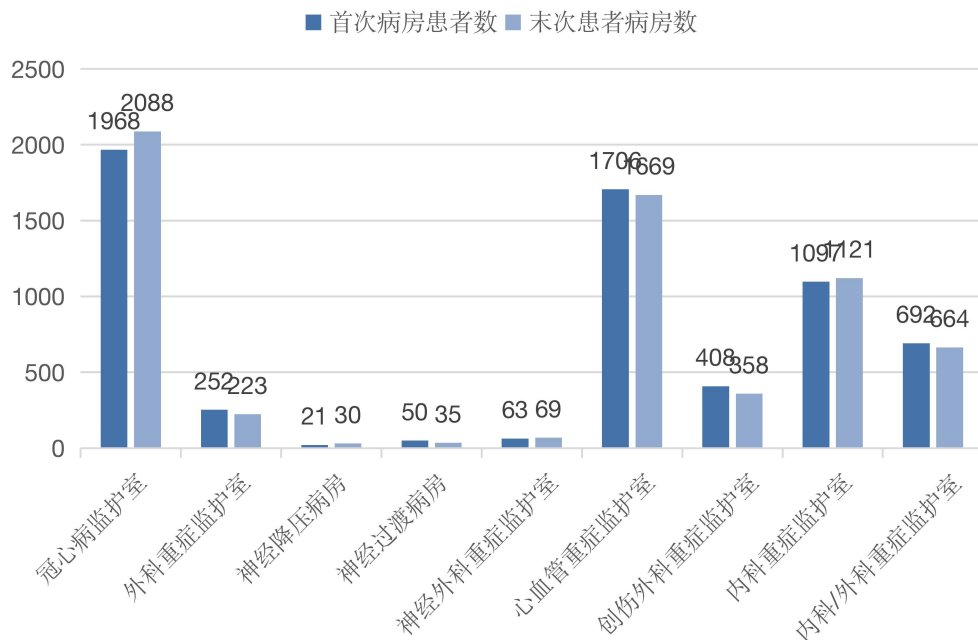


图 3.3 首末次病房类型分布

Figure 3.3 Distribution of first and last ward types

值得注意的是，人口统计信息中并未包含患者的身高体重。就体重而言，在患者的用药记录表、过程事件表、表格事件表中都存在记录，其中用药记录与过程事件中每一条目都包含患者体重，而表格事件中存在 4 项与体重相关的独立记录项，分别是每日体重、入院体重、饭后体重、入院体重 (lbs)。就身高而言，仅在“icu”中的表格事件表中存在，并通过英寸或厘米两类条目表示。

根据上述提及属性，绘制属性统计量表 3.2。其中 p 值代表样本统计结果存在差异时，由抽样误差所导致的概率大小。p 值越小则认为原假设发生概率越低。当 p 值小于 0.05 时，认为结果具有统计学差异；当 p 值小于 0.01 时为显著的统计学差异；当 p 值小于

0.001 时结果表现为显著的统计学差异。

表 3.2 属性统计量

Table 3.2 Attribute statistics

属性名	生存患者 (n=5667)	死亡患者 (n=590)	t 值	X ² 值	p 值
年龄, 均值 (4 分位数)	70 (62-81)	76 (68-85)	9.9628	-	<0.001
性别 (男), n (%)	3137 (55)	316 (53)	-	0.6264	0.4286
首次病房类型, n (%)					
CCU	1727 (30)	241 (40)	-	26.188	<0.001
SICU	370 (6)	38 (6)		0	1
Neuro stepdown	21 (0)	0 (0)	-	1.2257	0.2682
Neuro intermediate	48 (0)	2 (0)	-	1.1579	0.2818
Neuro SICU	48 (0)	15 (3)	-	13.755	<0.001
CVICU	1641 (29)	65 (11)	-	85.821	<0.001
TSICU	230 (4)	22 (3)	-	0.07713	0.7812
MICU	966 (17)	131 (22)	-	9.4768	<0.01
MICU/SICU	616 (11)	76 (13)	-	1.9981	0.1574
末次病房类型, n (%)					
CCU	1818 (32)	270 (46)	-	44.377	<0.001
SICU	329 (6)	29 (5)	-	0.6288	0.4277
Neuro stepdown	30 (1)	0 (0)	-	2.127	0.1447
Neuro intermediate	66 (1)	3 (0)	-	1.5508	0.2130
Neuro SICU	23 (1)	12 (2)	-	22.619	<0.001
CVICU	1618 (29)	51 (9)	-	107.25	<0.001
TSICU	207 (4)	16 (3)	-	1.116	0.2907
MICU	984 (17)	137 (23)	-	12.068	<0.001
MICU/SICU	592 (10)	72 (12)	-	1.5586	0.2118
婚姻状况, n (%)					
单身	1226 (22)	105 (18)		4.4723	<0.05
已婚	2606 (46)	270 (46)		0.0035	0.9521
其他	272 (5)	40 (7)		4.0134	<0.05
离婚	415 (7)	25 (4)		7.3183	<0.01
寡妇/鳏夫	1148 (20)	150 (25)		8.3627	<0.01

表 3.2 续

Table 3.2 Continued

属性名	生存患者组 (n=5667)	死亡患者组 (n=590)	t 值	X ² 值	p 值
人种, n (%)					
白种人	4078 (72)	431 (73)	-	0.2637	0.6075
其他	702 (12)	82 (14)	-	0.9792	0.3223
非裔美国人	576 (10)	51 (9)	-	1.2059	0.2721
亚洲人	126 (2)	16 (3)	-	0.3757	0.5399
西班牙裔/拉丁裔	170 (3)	10 (0)	-	2.8063	0.0938
印第安人/阿拉斯加原住民	15 (0)	0 (0)	-	0.6542	0.4185
入院来源, n (%)					
门诊手术转移*	3 (0)	1 (0)	-	-	0.3271
网站	102 (2)	4 (0)	-	3.3932	0.0654
专业护理机构转院	87 (2)	19 (3)	-	8.1277	<0.001
自行转诊/自主入院	116 (2)	18 (3)	-	2.1130	0.1460
医生转诊	1439 (25)	85 (14)	-	34.410	<0.001
信息不详	10 (1)	2 (0)	-	0.1327	0.7156
麻醉术后恢复*	3 (0)	1 (0)	-	-	0.3271
急诊入院	2293 (40)	270 (46)	-	5.9906	<0.05
诊所转诊	51 (1)	7 (1)	-	0.2165	0.6416
医院转出	1548 (27)	184 (31)	-	3.8076	0.051
保险类型, n (%)					
Medicare	3317 (59)	410 (70)	-	26.196	<0.001
Other	2119 (37)	170 (29)	-	16.581	<0.001
Medicaid	231 (4)	10 (1)	-	7.5520	<0.01
入院方式, n (%)					
欧盟观察*	3 (0)	0 (0)	-	-	0.7429
紧急入院	1325 (23)	146 (25)	-	0.4802	0.4883
择期入院	311 (6)	9 (1)	-	18.535	<0.001
手术日入院	417 (7)	7 (1)	-	31.258	<0.001
急诊科紧急情况	2596 (46)	309 (52)	-	8.9941	<0.01
观察入院	703 (12)	92 (16)	-	4.6135	<0.05
急诊科.	291 (5)	26 (4)		0.4474	0.5035
直接观察*	1 (0)	1 (0)		-	0.1797

*期望频数过小时使用 Fisher 精确检验

根据上表的统计性分析,可发现不同临床结局患者间在某特定属性划分下的差异性是否显著。如年龄、人种等不同属性划分下,均不存在统计学的显著性差异,而婚姻状况下的5种划分方式中的4种都具有显著性差异。该表的显著性结果将决定研究纳入的信息量。

3.2.3 事件表分析

患者记录纳入了三种类型事件,分别为过程事件、排出事件、表格事件。过程事件中共138项事件类型,总条目数64,448。根据事件类型又可划分为透析相关、药物使用、外周静脉插管相关、医学影像、重大事件记录、侵入式插管相关、生殖泌尿系统及胃肠道系统相关、细菌真菌培养相关等12项,药物使用类仅记录前列环素(Epoprostenol)的使用,该药物主要预防血栓形成。细菌真菌培养的记录仅表示做过培养,但无检验结果。表中记录频次最高项目依次为20号静脉插管、心电图检测、18号静脉插管、胸部X光、动脉插管、有创呼吸机、拔管、经胸超声心动图检测等过程。共6,255名患者有该表记录。

排出事件中共61项事件类型,条目数390,412,事件类型又可分为引流量及患者排出量两类。记录频次最高项为Foley导尿管中的排出液体量,共285,701条,而后是胸腔引流管。两类条目共占326,236条记录。其中Foley导管一般用于排尿或灌注药物等液体。胸腔引流管主要为胸部治疗提供帮助,排出胸腔积液等。该表中大部分记录项以导管种类命名。共6,168名患者有该表记录。

表格事件中记录事件共1,969项,条目数26,276,902。根据事件又可分为实验室检验、生命体征、体外膜肺氧合(extracorporeal membrane oxygenation, ECMO)配置、疼痛镇静管理相关、呼吸系统、消化道及泌尿生殖系统、常规记录、外周静脉插管相关、侵入式插管相关等39类。ICU设备相关的配置中在该表中唯一记录,如呼吸机配置,ECMO配置、透析仪配置等。该表还记录了与插管相关的现象评估,如是否出现插管部分感染、使用敷料等。记录频次最高项目依次为心率、呼吸率、血氧饱和度、心律、血压、体温等。所有患者都存在表格事件记录。

3.2.4 特殊条目分析

数据库中存在复合型条目,如格拉斯哥昏迷评分(glasgow coma scale, GCS)。格拉斯哥昏迷评分由三个检验条目类型构成,分别是睁眼反应、语言反应、运动反应,数据库相关记录字段见表3.3,括号内为数据库记录字段的英文说明。格拉斯哥昏迷评分通过度量三种条目的总分来评估患者的意识状态,因此被认为是复合型的。其总分值在3分~15分之间,临床上根据总分所在区间进行级别评测,分别是意识清醒(15分)、

轻度意识障碍（13~14 分）、中度意识障碍（9~12 分）、8 分以下为昏迷（3~8 分）。

表 3.3 GCS 字段

Table 3.3 GCS fields

睁眼反应	语言反应	运动反应
4 分：自动睁眼 (Spontaneously)	5 分：回答正确 (Oriented)	6 分：遵嘱动作 (Obeys commands)
3 分：呼唤睁眼 (To speech)	4 分：回答错误 (Confused)	5 分：刺痛定位 (Localize Pain)
2 分：刺痛睁眼 (To pain)	3 分：只能说话 (Inappropriate words)	4 分：刺痛躲避 (Flew-with-draws)
1 分：不睁眼 (None)	2 分：只能发音 (Incomprehensible sounds)	3 分：刺痛屈曲 (Abnormal Flexion)
-	1 分：无言语 (No response)	2 分：刺痛强直 (Abnormal extension)
-	-	1 分：无反应 (No response)

(*) 括号内为数据库中的英文字段名

数据库中还存在评估意义相似的条目类型。血糖作为常用的临床检测指标，库中分为全血血糖、血清血糖、手指测血糖三类条目。根据临床知识，手指端血糖的测量准确度较低，且餐后可能有较大浮动。血压在库中记录又分为侵入式与非侵入式两种，通常侵入式测量方式更为准确。体温有多种记录格式（见 3.3 节）。

3.3 数据清洗

本节将从缺失值处理、异常值处理、合并项处理分别阐述具体方法。临床中，界定异常值的最大最小范围较为困难，尤其是生化指标的可能范围，常需要人为判定。ICU 患者中的临床指标多为极度异常状况，但某些数值也存在明显记录异常问题，如患者心率记录值已超过 300，甚至存在 1,000 以上的数值。针对该类问题，使用两侧分位数替代明显异常值^[48]。

缺失值处理时，首先需要评判数据的缺失率，一般当缺失率达到 30%时将丢弃该特征。然而通过固定时间窗口提取临床特征的状态或大小时，受限于检验或记录的频次差异，使用较小的时间窗口将导致部分临床指标出现缺失，即该时间段内并未存在检验或记录。若将此种状态下特征的缺失率作为评判依据进行特征筛选，难免会损失大量有效

的检验记录信息。使用更大的时间窗口是可取的,但在此种模式下,算法的短时评估风险准确度也将随之下降。因此本研究使用填充方式处理缺失数据。缺失值填充时,常用方法包括最近值填充,均数填充、插值填充、众数填充以及机器学习模型填充。针对心率、血氧饱和度等数值型特征,最终使用线性插值填充法或均数填充法。针对心律、格拉斯哥昏迷评分等离散型特征,使用众数填充或最近值填充。具体而言,当患者记录数据中未存在某特征记录时,使用样本均数或众数,否则使用插值或最近值。

为了数据条目表示的一致性,将整合数据库中临床含义相近指标。其中侵入式与非侵入式血压若在同一时间窗口内存在数次记录,优先取最近测量值。体温、身高、体重均存在不同单位的记录,最终以摄氏度、厘米、千克的换算结果为准。血糖检验项包括血清血糖、全血血糖、手指末端血糖,虽然检测的精确度存在差异,仍考虑合并三种检验条目,若同一时间窗口内存在多类血糖记录,优先级为血清血糖>全血血糖>手指末端血糖。最终合并方式见表3.4。

表3.4 特征映射说明

Table 3.4 Description of the feature mapping

输入特征	映射特征	输入特征	映射特征	输入特征	映射特征
动脉收缩压	收缩压	动脉舒张压	舒张压	全血血糖	血糖
非侵入式收缩压	收缩压	非侵入舒张压	舒张压	手指末端血糖	血糖
动脉平均血压	平均血压	汇总呼吸率	呼吸率	血清血糖	血糖
非侵入式血压	平均血压	呼吸率	呼吸率	摄氏度体温	体温
入院体重 (kg)	体重	日常体重	体重	身高 (inch)	身高
入院体重 (lgs)	体重	饭后体重	体重	华氏度体温	体温
全血钠离子	血清钠离子	全血钾离子	血清钾离子	血清肌酐	肌酐

虽然部分临床指标的来源、检验目的、测量精度存在一定差异,但从减少计算稀疏性、简化特征工程以及数据规模来看,这种差异是可以在计算中被缩小的。

3.4 本章小结

本章节从多个角度分析了 MIMIC-IV 数据集的情况。并结合统计学、常用数据分析手段进行了可视化分析及数理统计分析。为了计算模型的输入合理性,特别对数据的异

常值、缺失值状况进行了处理，并对含义相近的临床指标进行了合并，便于后续的计算输入构造及数据建模。

第4章 基于 TFA-GRU 的心衰患者死亡风险预测模型

4.1 方法概述及数据预处理

电子病历中数据十分复杂，临床医师很难在短时间内发掘电子病历中的多样信息关联，可能忽略患者潜在风险。受 Che 等人^[28]启发，患者的死亡风险与临床干预模式存在一定关系，高死亡风险的患者需要更频繁的临床干预以维持生命体征，而相对低风险患者则具备更加稳定的干预治疗模式，通过可视化患者电子病历数据中每小时内产生的临床记录数据量，可发现部分患者的记录数变化存在一定规律（图 4.1）。图中横轴表示患者进入 ICU 后经过小时数，纵轴表示当前小时内电子病历的所有检验记录数。数据来自随机采样的 10 名患者的 24 小时数据，左侧为心衰院内死亡患者，右侧为生存患者。生存患者呈现的记录数变化更具规律性，而院内死亡的患者记录数很难发现出规律的片段周期。这种现象可能来自患者生理状况差异，也表明生存患者的干预较持续且稳定。而心衰院内死亡患者短时间内体征变化大，需要更多临床干预。值得注意的是，图中后三行患者虽然是生存患者，但其记录数的变化并不太稳定，可理解为该类患者在院期间的生理体征并不太稳定，虽然没有在心衰住院期间死亡，可能仍存在较高死亡风险。综上，本章节将提出 TFA-GRU (Temporal Frequency-Aware Gated Recurrent Unit) 深度模型，以学习此类患者临床记录模式。

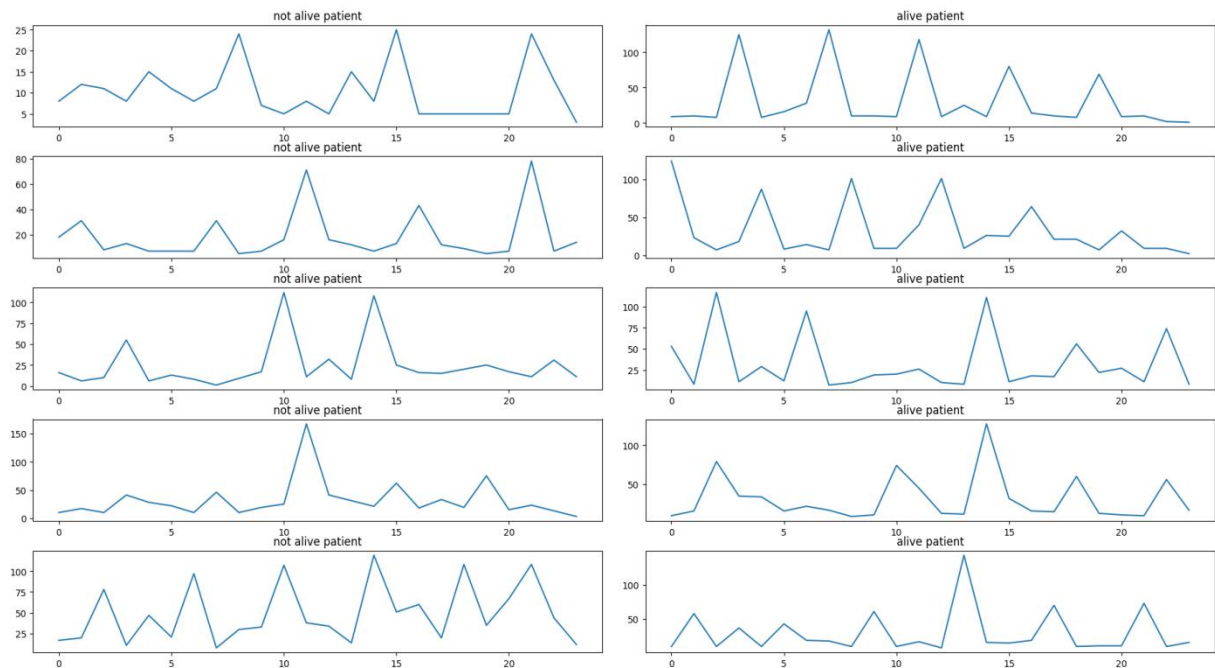


图 4.1 不同临床结局患者的小时记录数折线图

Figure 4.1 Line chart of hourly records for patients with different clinical outcomes

在 MIMIC-IV 数据中, 血氧饱和度、心率等是按每 1 小时进行记录。因此, 本研究以每 1 小时为时间窗口提取临床特征。通过文献分析^[20,29,49]、专家意见, 算法将纳入血压、血氧饱和度、心率、心律、ICU 行为记录、血常规、电解质、格拉斯哥昏迷评分等临床量表、手术过程等多源数据共 74 项, 见附录 A-2。

输入数据分为两部分。一部分称为静态数据, 即不随时间变化的数据, 如种族、性别、诊断等。第二部分称为动态数据, 即短时间内容易发生变化的数据, 如患者的血压、心率、血氧浓度、钠钾离子浓度等。数据预处理时以每小时内的记录数据进行汇总统计, 记录每小时内的总临床记录数以及距离 ICU 入院后的时间差, 该时间差以小时为单位。由于每小时所形成的输入向量是固定维度的, 将导致某些低记录频率或低检验频率指标缺失。针对上述状况, 按 3.3 节方法进行值填充, 为区别不同数值在当前时间窗口内的记录状态, 引入指示向量。指示向量与输入向量维度相同, 1 表示对应特征在当前窗口内实际记录, 0 表示当前特征是填充值。具体针对离散型特征, 记录类型对应的指示向量位为 1, 其余为 0。针对数值型特征, 存在记录时为 1, 否则为 0。

由于临床特征的方差、极差均较大, 对于数值型特征的规范化, 在对比使用最大最小标准化 (Min-max) 和中心标准化 (Z-score) 后选用中心标准化方式的效果更加, 最大最小标准化方式将使得数据值分布出现长尾效应。标准化时采用的均值方差为总体样本对应的统计量。所有的类别型指标用独热编码表示, 过程信息使用二值向量表示。输入及指示向量构造方法见图 4.2。

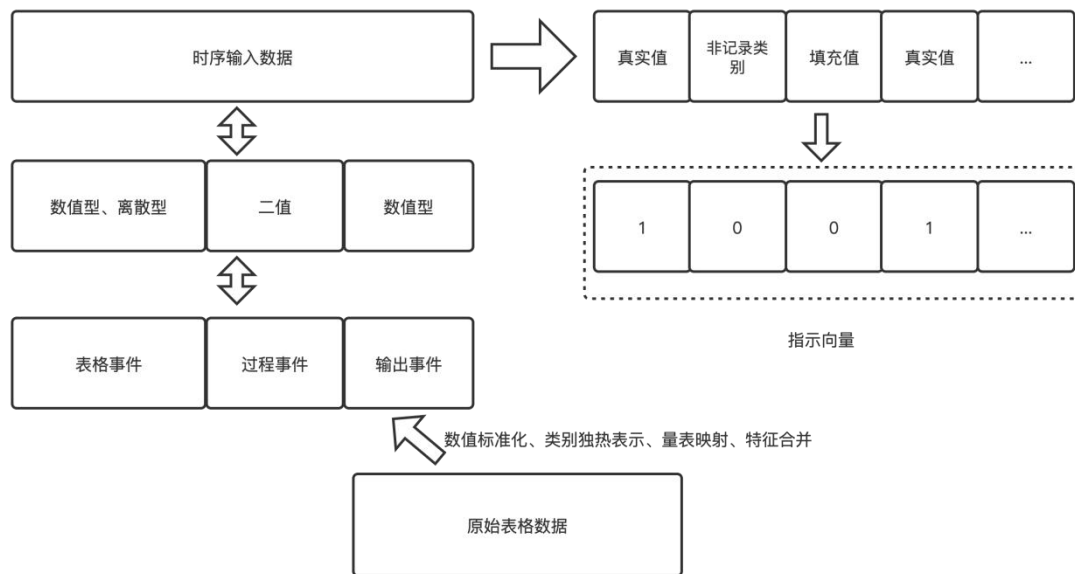


图 4.2 输入数据及指示向量构造

Figure 4.2 Constructions of input data and instruct vector

4.2 模型结构及细节

TFA-GRU 为充分学习患者电子病历中的多源异构的复杂信息，将模型分为 4 个模块，分别是时频编码模块、时序数据编码模块、死亡风险预测模块、编码诊断模块组成。时频编码模块主要对入院时间，临床干预频次、临床特征联合学习，拟学习到不同患者临床特征间的时频关联以及隐藏风险模式。首先，将患者每小时内的各项数据填充，并通过指示向量标识状态，而后将指示向量乘以权重向量。另外，将患者 24 小时内记录的历史频次经过 softmax 层归一化转换，为了避免 softmax 函数对于极大数较敏感的问题，添加温度参数^[50]适当拉平频次分布，温度参数越大，分布越平缓。归一化后的频次向量及入院时间间隔向量均通过不同的线性层投影到高维后相加，并与加权后的特征向量拼接，最后通过逐时间步的线性层融合时间、频次、临床特征信息，得到时频感知向量（temporal frequency-aware vector, TFA-vector），时频编码模块见图 4.3。

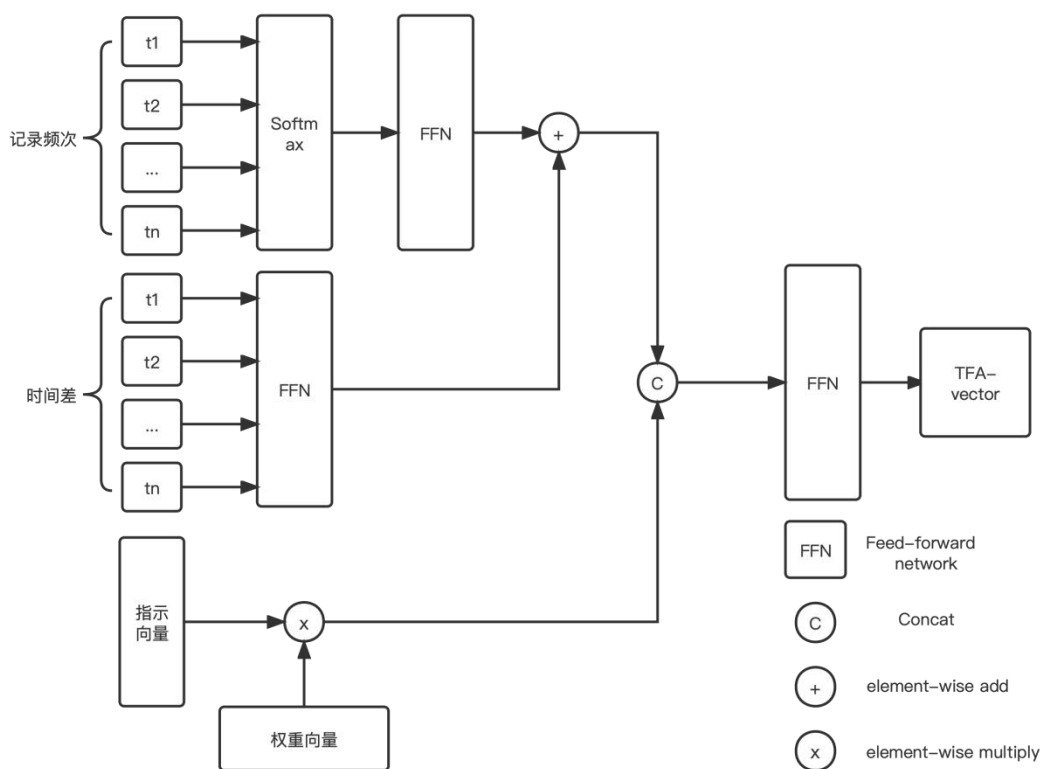


图 4.3 时频编码模块

Figure 4.3 Temporal frequency-aware encoding module

定义患者电子病历中小时总记录数向量为 $c=\{c_1, c_2, \dots, c_n\}$ ， $c_i \in \mathbb{N}$ 。入院后的时间间隔向量为 $t=\{t_1, t_2, \dots, t_n\}$ ， $t_i \in \mathbb{N}$ ，则时频编码模块的输出计算公式如下：

$$v_{tc} = \text{FFN}_3 \left(\text{CONCAT} \left(\text{FFN}_1 \left(\text{Softmax} \left(\frac{c}{\tau} \right) \right) + \text{FFN}_2(t), \omega \cdot v_{instruct} \right) \right) \quad (4-1)$$

式中：FFN —— 前馈神经网络； τ —— 温度系数； $v_{instruct}$ —— 指示向量； ω —— 特征加权向量；

v_{tc} —— TFA-vector

时序数据编码模块主要针对时间序列数据建模，因此仅接受动态数据进行参数学习。输入包含时频编码模块输出的时频向量及预处理后的患者多源数据。时序数据编码模块由多层 Bi-GRU (bidirectional gated recurrent unit) 构成，以获得患者时间维度的编码特征，该模块的最终输出由末时间步中最后一层的隐层状态决定。

编码诊断模块是经过 word2vec 方式预训练的嵌入层以及加权聚合处理。该模块训练预先独立于其余模块，需先基于整体诊断数据通过 skip-gram 进行诊断向量的学习。在实际训练中将预训练好的嵌入层权重引入，患者间不同长度的诊断信息通过嵌入层后经平均聚合并结合诊断优先级输出表示向量，模块参数参与训练。

死亡风险预测模块的输入包含时序数据编码模块输出、患者静态数据、诊断编码模块输出并输出患者的死亡风险。该模块中引入的患者静态信息为 3.2.2 节中差异呈显著性的属性。其中 ICU 入院类型、首末次 ICU 监护室类型息使用独热向量，年龄使用标准化后的数值表示。特征融合模块中首先将所有输入拼接的向量经过一层 Layer norm 正则化层，然后经过多层的前馈神经网络及非线性函数，最终通过 sigmoid 激活函数输出患者的死亡风险，公式如下：

$$\text{output}_{\text{risk}} = \text{Sigmoid}(\text{FFN}(\text{CONCAT}(v_{\text{static}}, v_{\text{GRU}}, v_{\text{diag}}))) \quad (4-2)$$

式中： v_{static} —— 静态信息向量； v_{GRU} —— 时序数据模块输出向量； v_{diag} —— 编码诊断模块输出； $\text{output}_{\text{risk}}$ —— 死亡风险

TFA-GRU 结构见图 4.4。 $v1 \sim vn$ 为 TFA-vector 对应的不同时间窗口向量。 $t1 \sim tn$ 为数据预处理后的各时间步患者表示向量。

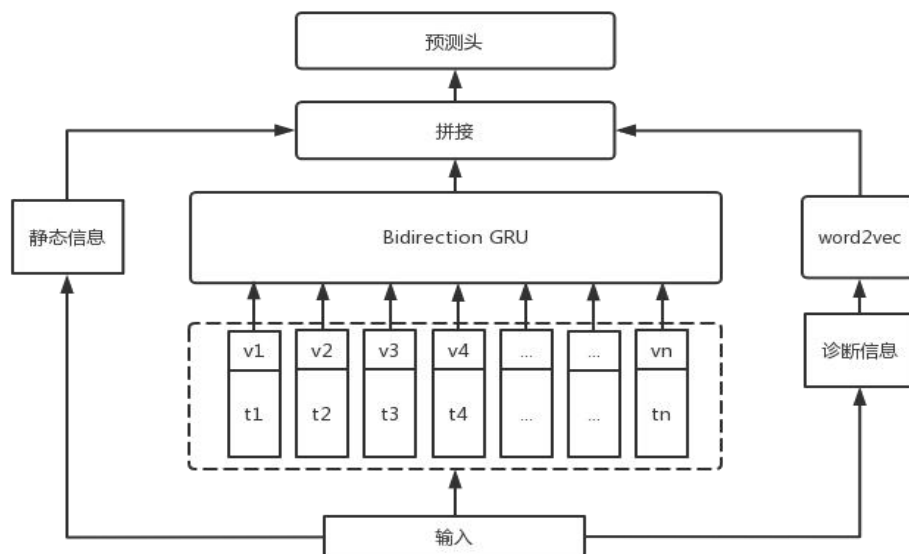


图 4.4 TFA-GRU 整体结构

Figure 4.4 Structure of TFA-GRU

TFA-GRU 模型面向 ICU 内患者的短时死亡风险评估, 定义 6,257 名患者中 590 名心衰住院期间死亡患者标签为 1, 否则为 0。训练时以 FocalLoss^[51]作为损失函数, 其弥补了交叉熵损失在正负样本不均衡时损失函数将倾向样本量多的类别问题, 并引入可调节参数 γ 对难易样本区分学习。在模型训练过程中, 根据验证集的评估损失使用早停方法, 优化器对比使用 SGDM (Stochastic Gradient Descent with Momentum)。

$$\text{FocalLoss}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (4-3)$$

式中: p_t —— 模型输出概率; γ —— 可调节系数; α_t —— 调节权重

算法 1 TFA-GRU 训练过程

输入: 样本集 $S = \{(t^i, c^i, v^{i_{ehr}}, v^{i_{insturct}}, y^i), \dots\}$, 其中 $v^{i_{ehr}}$ 为电子病历记录向量, t^i 为患者输入序列的时间差; c^i 为对应时间内的总记录数; $v^{i_{insturct}}$ 为临床记录的填充指示向量; y^i 为患者是否出院死亡标签; TFA-GRU 模型 f ; 最大训练周期 $T_{\max}$; 早停数 patience ; 训练批大小 bs_size ; 训练集大小 N ; 历史损失记录 pre_loss

输出: 患者的短期死亡风险 $\text{output}_{\text{risk}}$

```

1:  $\text{pre\_loss} = 1e8$ 
2: 划分样本集为训练集  $S_{\text{train}}$ 、验证集  $S_{\text{eval}}$ 、测试集  $S_{\text{test}}$ 
3: for 0  $\rightarrow T_{\max}-1$  do
4:   for 0  $\rightarrow N // \text{bs\_size}$  do
5:     从训练集合  $S_{\text{train}}$  抽取  $\text{bs\_size}$  的样本  $S'$ 
6:     计算  $S'$  的损失  $\text{loss}$ , 并使用 SGDM 进行梯度下降优化模型  $f$ 
7:   end for
8:   计算  $f$  在  $S_{\text{eval}}$  上的损失  $\text{eval\_loss}$ 
9:   if  $\text{eval\_loss} \geq \text{pre\_loss}$  then  $\text{cur\_patience} += 1$ 
10:    else  $\text{cur\_patience} = 0$ 
11:  end if
12:   $\text{pre\_loss} = \text{eval\_loss}$ 
13:  if  $\text{patience} == \text{cur\_patience}$  then 停止训练
14:    else 继续训练
15:  end if
16: end for
17: end for

```

4.3 实验结果

模型训练集、验证集、测试集比例为 0.7:0.15:0.15。参数配置上, Bi-GRU 的层数为 4, 隐层大小为 256, dropout 为 0.4, skip-gram 窗口大小 2, 输出向量大小为 32。为验证 TFA-vector 引入的有效性, 引入表 4.1 对比不同维度的 TFA-vector 对最终结果的影响。

从表 4.1 发现, 未使用 TFA-vector 的 GRU 模型表现性能较低, 而使用 TFA-vector 后的模型性能普遍提升。过于高维的 TFA-vector 可能扰乱原始电子病历输入分布, 过于低维的时频特征对门控单元的学习起到较少的作用。

表 4.1 模型训练结果

Table 4.1 Training results

类型	TFA-V	TFA-V dim	AUROC	AUPRC
GRU	F	-	0.938	0.802
TFA-GRU ₁₂₈	T	128	0.944	0.817
TFA-GRU ₄₈	T	48	0.946	0.828
TFA-GRU ₆₄	T	64	0.938	0.798
TFA-GRU ₃₂	T	32	0.943	0.818
TFA-GRU ₁₆	T	16	0.939	0.805
TFA-GRU ₈	T	8	0.943	0.811
TFA-GRU ₄	T	4	0.942	0.813
TFA-GRU ₂₅₆	T	256	0.945	0.822

TFA-V 列代表输入是否有时频编码模块向量, TFA-V dim 为时频编码模块输出的向量维度

从表 4.1 中发现不同模型间的性能差异并不大, 尤其是 AUROC, 但从 AUPRC 看 TFA-GRU₄₈ 与 TFA-GRU₂₅₆ 都有相对较高的性能表现, 目前来看 TFA-GRU₄₈ 是最优的死亡风险预测模型。为验证模型间的临床效益评估, 绘制不同模型间的 DCA 曲线汇总图 4.5。

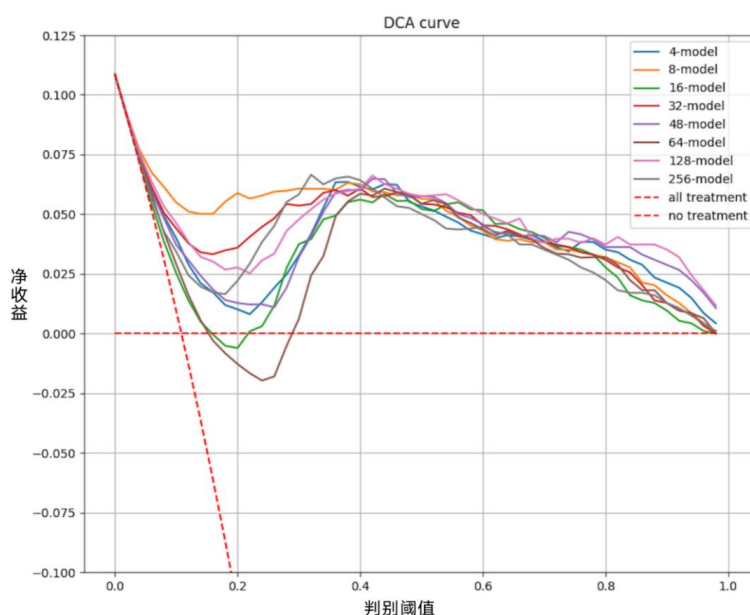


图 4.5 不同参数下模型的 DCA 曲线

Figure 4.5 DCA curves of the model under different parameters

在 DCA 曲线中, y 轴值越大代表对应判别阈值下的模型具有愈高的临床净收益, 模型更佳。从图 4.5 可知, 不同时频向量模型在中部判别阈值范围下的净收益是相近的, 但在较高阈值状态下, TFA-GRU₁₂₈ 和 TFA-GRU₄₈ 网络表现性能更好。而在低阈值下同样存在较大差异, TFA-GRU₈ 模型的净收益更高。从整体判别阈值空间下, 净收益抖动较小的也是 TFA-GRU₈。由公式 (2-7) 易知, 净收益仅与当前判别阈值下的真正例与假正例相关, 可理解为净收益高的模型, 其对正例的辨别度更高。另由公式 (4-1) 可知目标损失函数可关注难样本的学习, 在此函数训练下将使得模型的预测更加两极化, 难样本也将越来越少, 取而代之的是相对低或高死亡风险的患者预测, 这也符合中部判别阈值下模型预测相近, 两端会出现性能差异的现象。综上, DCA 曲线与 AUROC 或 AUPRC 的度量方式有所不同, 所评估出的最佳模型也是不同的, 这说明了在临床中选择算法的依据与传统机器学习评价指标的意义并不相同, 不能仅依靠机器学习指标评估算法有效性。

TFA-GRU 模型的死亡风险由多个信息共同预测, 为探究每个模块信息对于模型效果的影响, 分别将静态信息 (No static)、诊断信息移除 (No diag), 探究最终性能效果 (图 4.6)。从图中可知, 在未纳入静态信息的状态下, 模型仅通过诊断信息和时序信息就达到了更优的表现性能。可知, 静态信息的引入反而降低了模型的表现能力。这或许与表 3.2 的统计量结果存在一定关系, 如首末次病房类型中, 存在并不具备显著性区分能力的类别样本, 引入反而干扰了模型的死亡风险辨别能力。

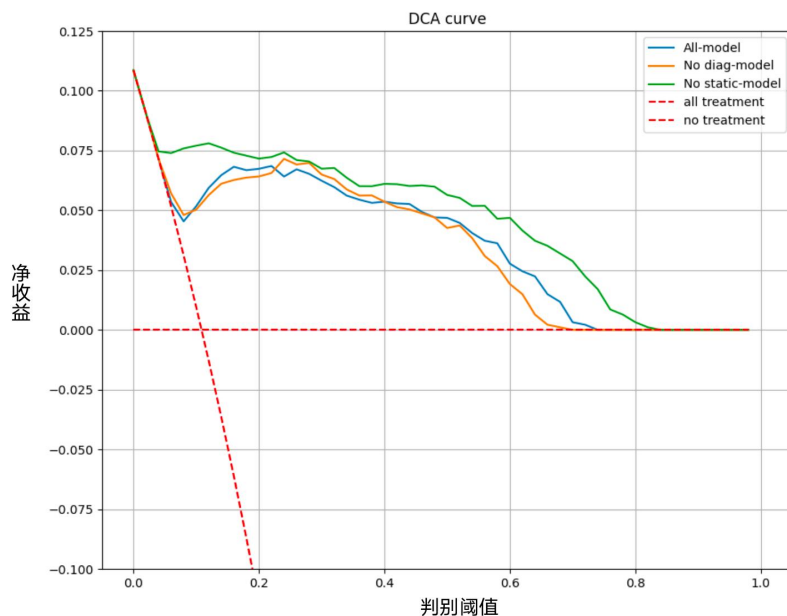


图 4.6 信息影响对比

Figure 4.6 Information impact comparison

为评估 TFA-GRU 的实际临床应用价值及有效性, 将模型对于心衰入院死亡及非死亡样本组的预测风险值进行可视化展示 (图 4.7)。可知死亡组及生存组的死亡风险预

测概率分布存在显著差异。死亡风险组中，平均预测风险高达 0.8，而非死亡组仅 0.1 左右，且非死亡组的预测概率几乎在 0.3 以下，或可将其作为 TFA-GRU 对于高风险死亡患者的应用判别阈值。

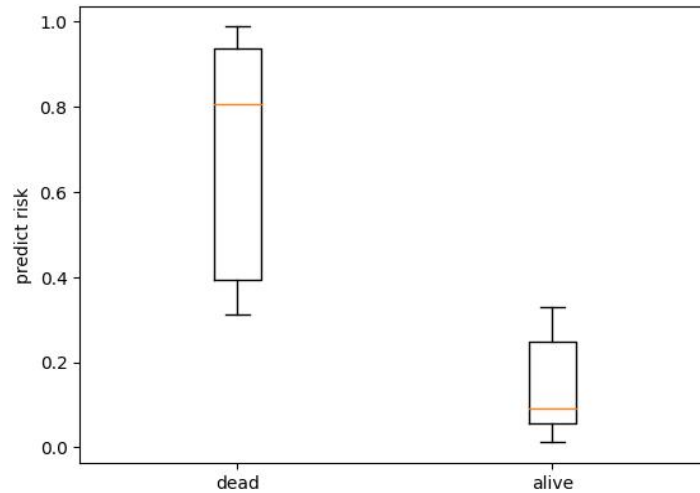


图 4.7 死亡组及生存组的模型预测风险值

Figure 4.7 Models of death and survival predicted risk

4.4 解释性分析

本节将通过 DeepSHAP 方法评估死亡风险与临床特征之间的关联，并结合可视化及临床先验知识分析结果的可信度。

首先从数据集中随机抽样 50 名患者，其中死亡与非死亡组的人数一致，通过 DeepSHAP 框架计算样本的 shapley 值，并绘制出采样样本汇总后与死亡风险高相关的前 20 项因素（图 4.8）。其中纵轴标签前缀 num 表示该样本是数值型特征，否则是 Richmond-RAS 是 Richmond 躁动-镇静评分。据图 4.8 可知，对死亡风险具有较大影响的临床特征依次是血氧饱和度（O2 saturation pulseoxymetry）、收缩压（Arterial Blood Pressure systolic）、舒张压（Arterial Blood Pressure diastolic）、心率（Heart Rate）、体温（Temperature Celsius）、阴离子间隙（Anion gap）、血清氯离子浓度（Chloride serum）、白细胞计数（WBC）、血红蛋白（Hemoglobin）、血清碳酸氢根浓度（HCO3 serum）、磷化物（Phosphorous）、血清钠离子（Sodium serum）、格拉斯哥昏迷评分（GCS）、窦性心律（Sinus Rhythm）、钾离子（Potassium）、乳酸（Lactic Acid）、尿素氮（BUN）、呼吸末正压（positive end-expiratory pressure, PEEP set）。

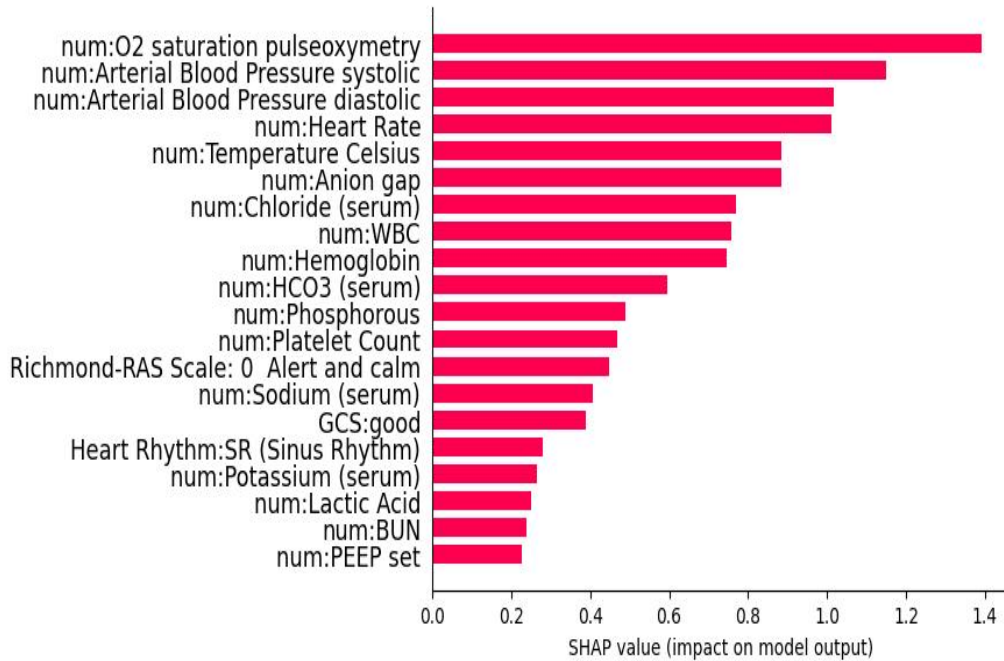


图 4.8 前 20 项死亡风险高相关因素 (绝对值)

Figure 4.8 Top 20 factors associated with risk of death (absolute)

图 4.8 说明了在抽样患者内各临床特征对于整体预测结果的绝对贡献水平，忽略了正负相关性。为了探究每个临床特征对于患者死亡风险的正负贡献关系，逐一绘制抽样患者的平均 shapley 值（图 4.9）。图 4.9 中水平方向代表特征对于死亡风险预测的贡献程度，向左表示该特征降低了死亡风险，反之则增加死亡风险，而颜色代表该患者特征的 24 小时内平均大小，散点越红代表值越大，越蓝代表值越小。不难看出患者平均心率越低，其死亡风险逾大，同理血氧饱和度、血红蛋白浓度、磷酸盐含量、阴离子间隙也与死亡风险呈现负相关趋势。在图中尿素氮平均水平与死亡风险呈现出正相关关系，印证了刘^[52]等人的观点，认为血尿素氮或可作为评估心衰患者病情程度的重要指标。氯主要以氯离子形式存在于组织与体液中，对人体渗透压和酸碱平衡调节具有重要作用。在研究中发现氯离子水平与心衰患者的死亡风险呈现负相关关系，即血清氯离子浓度降低会增加心衰患者的死亡风险，也与安等^[53]研究结果一致。华等^[54]认为心肌梗死患者中，白细胞计数增加是心衰发生及死亡率增加的预测指标。在本研究结果中，白细胞计数与患者死亡率存在正相关趋势，也与其研究结果一致。对于非数值型的 Richmond 躁动-镇静评分、格拉斯哥昏迷评分而言，值越大表示该评估项符合次数越多，可见患者在平静、意识清醒状态下的风险更低。窦性心律是正常心跳的节律，其状态的持续也将表明患者的死亡风险越低。

值得注意的是，在某些患者中乳酸、动脉血二氧化碳分压的升高对患者死亡风险升高具有较大贡献，但相似数值下的其余患者却对死亡风险升高具有较低的贡献。可能是患者死亡风险构成往往是复杂且很难发掘的，仅从单一临床指标的高低程度评估患者风

险可能不够准确。

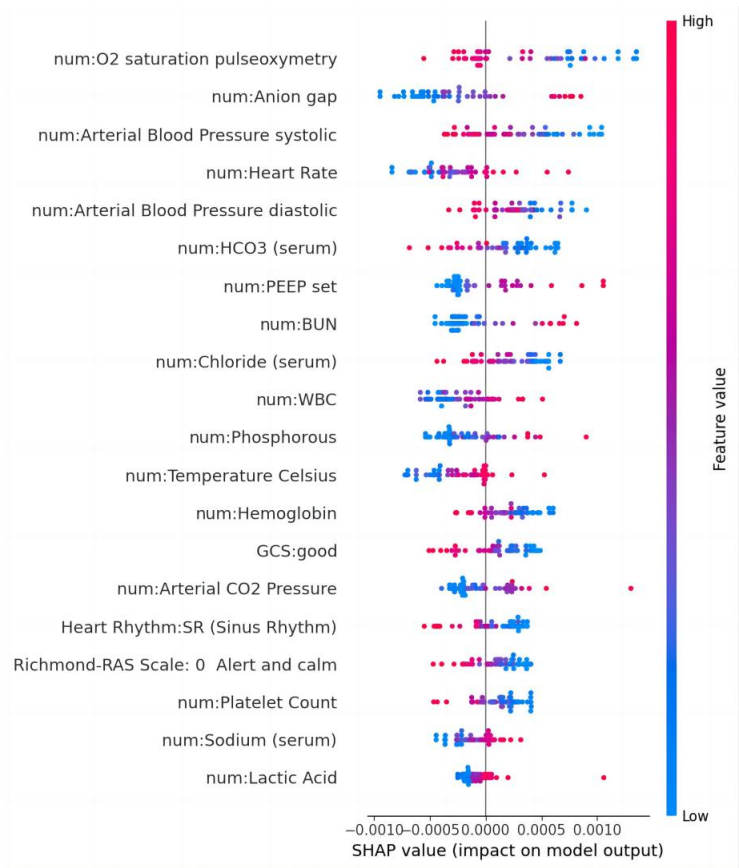


图 4.9 前 20 项死亡风险相关性临床特征贡献（平均）

Figure 4.9 Contribution of top 20 factors associated with risk of death (mean)

4.5 本章小结

本章节利用电子病历中的临床干预变化、诊断、人口统计、临床监护等多源信息，首次提出 TFA-GRU 网络结构，结果表明算法能准确预测心衰患者的死亡风险，性能可达 0.946 AUROC，0.828 AUPRC。另外，对于预测指标的解释性分析，也与其他研究者的研究结果一致。本研究发现较高的白细胞数以及较低的动脉血二氧化碳分压等临床特征与死亡风险呈现较高的相关性。

第5章 基于 Transformer 的心衰患者死亡风险预训练模型

5.1 方法概述及数据预处理

患者的诊断信息可看作是体征、症状、病因等的评估数据。每次就诊时，患者的电子病历中都将记录相关的评估信息。Transformer 作为一种能挖掘时序数据关联的网络结构，能基于历史诊断信息发掘患者各评估信息之间的关联，从而提高对患者的表示学习能力，发掘隐藏的风险。基于历史诊断的患者表示向量有利于下游死亡风险任务预测性能的提高。同时，MIMIC-IV 中的诊断信息具有缺失率低、格式统一的特性，其数据质量是相对高的。因此本章节将提出基于大规模患者诊断数据的心衰患者死亡风险预测预训练模型 SBFHF (Slim BERT For Heart Failure)。

首先将 MIMIC-IV 中的患者信息作为预训练语料库，提取患者诊断表中所有患者的临床诊断数据，并根据如下条件筛选数据，1) 去除患者诊断数小于 5 的患者数据，2) 去除年龄小于 18 岁的患者。由于心衰是不可逆的过程，定义心衰诊断患者为历史住院过程中产生过心衰诊断，非心衰患者为从未包含任何心衰诊断。具体数据提取流程见图 5.1。

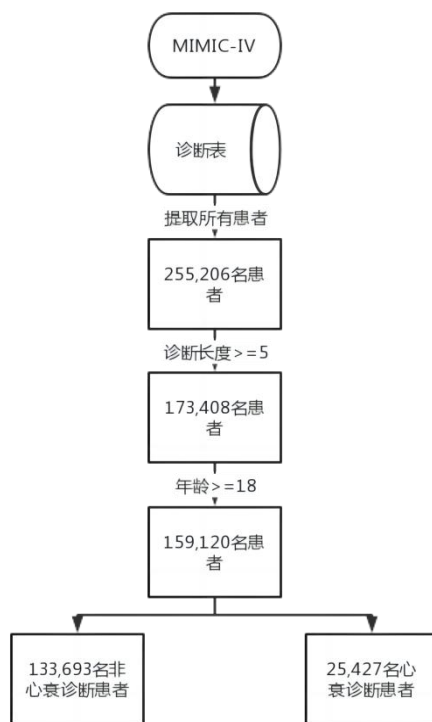


图 5.1 预训练数据提取

Figure 5.1 Pre-training data extraction

最终筛选出共 159,120 名患者，住院记录 422,749 条，相关诊断 4,946,806 条。其中心衰诊断患者 133,693 人，非心衰诊断患者 25,427 人，IR 为 5.258。预训练阶段的诊断数据相关统计见表 5.1。

表 5.1 预训练阶段的诊断数据统计

Table 5.1 Statistics of diagnostic data during the pre-training

诊断序列长度	值	就诊次数	值
均值	31.08	均值	2.656
最小值	5.000	最小值	1.000
最大值	2,087	最大值	238.0
标准差	53.52	标准差	3.737
中位数	15.00	中位数	2.000
90%	66	90%	5

由表 5.1 可知患者间的诊断序列长度及就诊次数存在较大差异，如何对该类分布不均衡的数据进行建模十分关键。目前，基于 Transformer 的电子病历诊断信息提取已有一些相关工作^[33,54,55,57,58]，主要区别在与嵌入标记的种类以及代理任务定义。其中 CEHR-BERT^[54]引入人工时间标记 (artificial time token, ATT) 区分不同就诊的间隔时间，结合实际就诊时间、患者年龄、临床概念共四种嵌入标记共同作为模型输入。代理任务包括 MLM 以及编码相关类型预测 (visit type prediction, VTP)，VTP 即预测临床编码是门诊、住院、急诊类型等种类。Med-BERT^[33]中引入了三种类型嵌入标记，分别是临床概念、临床概念优先级、就诊顺序。在代理任务选择上，使用了 MLM 以及延长住院时间 (Prolonged LOS) 预测任务，即预测患者住院延长是否超过 7 天。Hi-BEHRT^[55]中使用患者年龄、临床概念、就诊分隔符、就诊顺序四种嵌入标记。网络结构与传统 Transformer 编码器不同，底层嵌入标记聚合特征通过指定滑动窗口大小的 Transformer 编码为片段特征，并构造层级堆叠的 Transformer 编码器汇聚片段特征信息。代理任务上，通过对比学习方法 BYOL^[56] (Bootstrap Your Own Latent) 改进了传统 MLM 的任务训练方式，具体将 MLM 中的遮罩或替换为噪声视为一种数据增强方法，分别在临床概念及片段特征上进行增强并最小化目标网络与在线网络的输出来指导网络更新。BEHRT^[57]中引入四种嵌入标记，分别为年龄、临床概念、就诊顺序、就诊分隔符。将预测某就诊的临床概念及 MLM 共同作为代理任务。Rao 等^[58]提出了改进的 BEHRT 结构，将输入的标记替换为就诊时间、年龄、临床概念三类构成，预训练代理任务仅纳入了 MLM，下游任务仅针对半年内事件型心衰的患病预测。5 种预训练模型对比见表 5.2。

表 5.2 不同预训练模型对比

Table 5.2 Comparison of different pretrained models

模型	输入标记类型	临床概念种类	代理任务	预训练患者规模
CEHR-BERT	就诊时间、年龄、临床概念、就诊间隔	诊断、药物、过程	MLM+VTP	2,400,000
Med-BERT	就诊顺序、临床概念、临床概念优先级	诊断	MLM+prolonged LOS	28,490,650
Hi-BEHRT	年龄、临床概念、就诊分隔符、就诊顺序	诊断、药物、过程、生理体征	MLM with BYOL	2,844,733
BEHRT	年龄、临床概念、就诊顺序、就诊分隔符	诊断	MLM+预测下次临床概念	1,609,024
改进的 BEHRT	年龄、就诊时间、临床概念	诊断、药物	MLM	1,609,024

上述模型主要聚焦患者电子病例的表示学习，对心衰患者的下游预测任务的针对性不高，且使用的数据集规模也是百万级以上，虽然 Rao 等工作聚焦于事件型心衰患者，且达到了不错的性能，但需要患者存在 10 次以上就诊记录，对数据具有极高要求。另外，其训练的权重也是并未开源的。

在预训练模型范式中，代理任务的设计、数据规模是预训练模型成功的关键。本章基于上述预训练模型基础，提出 SBFHF 模型架构，改进的 MLM 任务以及针对心衰患者的判别代理任务。改进的 MLM 任务一定程度上弥补数据集相对较小的问题，而判别代理任务使得模型能更好的识别心衰患者特点。模型结构上，仍然堆叠 Transformer 编码器结构，将患者历史诊断、患者年龄、就诊顺序、诊断优先级四种标记构成模型输入。SBFHF 仅基于历史诊断及常规信息，相较于前述章节模型，不需要患者院内短时间的监护数据及相关设备支持，具备更高的易用性。

数据预处理时，为防止相近临床语义的不同版本编码出现在序列的不同位置，干扰模型收敛问题，首先将 ICD-9 编码转化为 ICD-10 (level-4)。处理后共得到 ICD-10 编码共 1736 种。并绘制转换后 ICD-10 编码的出现频次（图 5.2）。通过图 5.2 的分析，为防止数据集相对较小，训练语料不足引起的欠拟合问题。将出现 ≤ 5 次的 173 个诊断编码视为罕见类型用[RARE]表示，定义临床概念词表大小为 1,563。

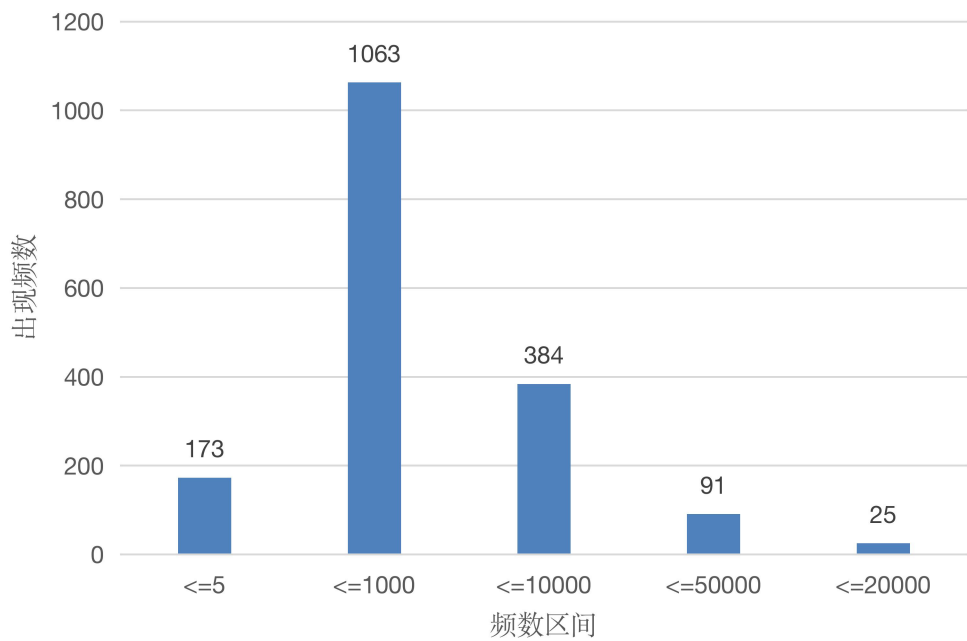


图 5.2 诊断编码出现频次

Figure 5.2 Frequency of occurrence of diagnostic codes

本章节研究目标患者与前述章节研究患者一致，并额外增加心衰住院之前的所有住院诊断信息，共得到 6,257 名患者的 23,037 次住院信息，以及 388,680 条诊断数据。数据诊断序列特性见表 5.3。

表 5.3 目标数据诊断数据统计

Table 5.3 Diagnostic data statistics for the target data

诊断序列长度	大小	就诊次数	大小
均值	62.11	均值	3.681
最小值	4.000	最小值	1.000
最大值	1989	最大值	136.0
标准差	95.17	标准差	5.470
中位数	31.00	中位数	2.000
90%	135.0	90%	8.000

根据表 5.3 及表 5.1 对比可知，目标数据集的诊断序列相比预训练数据而言整体更长，且仍然存在极大序列长度的问题。考虑到数据量规模及模型大小，针对较长序列，以目标数据的 90%值作为诊断序列长度可处理的最大值。超过将截取时间较近的诊断序列。

由于患者数据经去标识化处理，患者信息表中记录的年龄及入院时间仅是首次入院信息，因此只能通过患者的就诊时间以及最早就诊时间的差值推算患者年龄。如患者的

历史诊断记录发生在 2015-9-28 日，而患者信息表记录的最早入院时间是 2014-8-20 日，且当时年龄为 79，那么根据时间偏移可得到患者目前的年龄在 80 岁左右。在年龄标记的粒度选择上，绘制预训练数据的年龄分布（图 5.3）。由图 5.3 可知预训练数据中，患者年龄分布在 18~91 岁之间，其中 91 是所有大于 89 岁患者的映射年龄。整体来看并未出现过于稀疏的年龄片段，但数据更多集中在 60 岁左右。

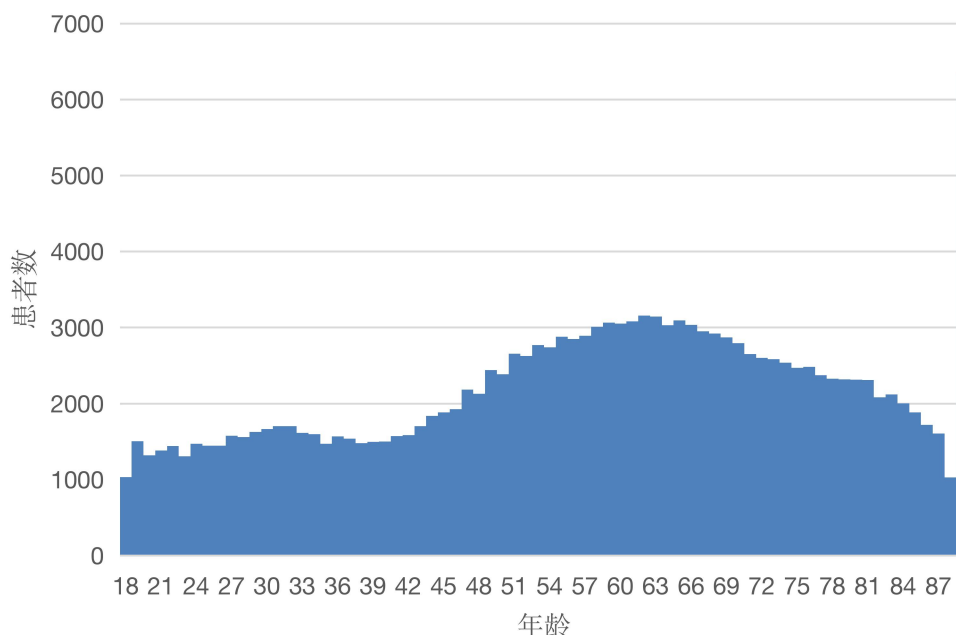


图 5.3 预训练数据年龄分布图

Figure 5.3 Histogram of age distribution of pre-training data

5.2 模型结构及细节

与已有预训练模型一致^[33,54,55,57,58]，SBFHF 仍使用基于 Transformer 的编码器结构，其提取患者电子病历特征的有效性在前述方法中得到了证明。另外，为构造更合理的输入诊断信息序列，SBFHF 考虑纳入患者就诊顺序、临床概念（ICD）、就诊时年龄以及 ICD 的优先级别。上述四项输入类型能带来单次就诊内的位置信息以及多次就诊间的位置信息，年龄也是常见的患者死亡风险高相关特征。

为将 SBFHF 更加适配于心衰患者群体，考虑更改传统掩码语言模型的代理任务以及训练方式。传统 MLM 任务常通过随机选择遮罩标记方式进行，可能导致某些词源标记得不到充分训练。因此算法增加对心衰相关诊断编码的遮罩或替换，同时适当平衡非心衰诊断编码的 MLM 任务参与度，将改进处理的 MLM 视为模型预训练代理任务一。该方法在数据规模相对小的前提下，不但能提高心衰诊断及其并发症互相感知能力，同时也能增加模型对不同心衰诊断的区分能力。另外将产生过心衰诊断与未包含心衰诊断

的患者进行区分，将判断患者是否有过心衰诊断作为代理任务二。心衰患者的判别任务可理解为在预训练集合中，给予心衰患者更大的关注度，并让模型尽量感知心衰患者与其余患者的异同，该任务同样可提高模型对患者诊断序列的整体总结能力。

在模型的输入构建方面，词表除包含诊断编码外，还包括[MASK]、[RARE]、[CLS]、[PAD]四种特殊标记。在预训练阶段中，[CLS]对应的序列输出用于判断是否有心衰诊断，微调阶段用于心衰患者的死亡风险预测。[RARE]表示数据集中定义的出现频次较少的诊断编码，也可用于表示推理时遇到未包含的诊断编码。[MASK]表示 MLM 任务中遮罩后的占位输入标记。针对患者间不同诊断长度序列问题，通过[PAD]占位符组成形状相同的批数据，提升训练速度。

SBFHF 的训练分为三个阶段，第一阶段为诊断编码及患者其他数据的表示嵌入处理过程，第二阶段为联合训练 MLM 任务及心衰患者识别任务，提高模型编码器提取特征能力，第三阶段则基于上一阶段的训练权重，在目标数据上进行整体网络权重微调以识别心衰患者的死亡风险。第三阶段不再使用掩码语言任务，输入仅包含常规诊断编码、[RARE]以及[PAD]占位符。

在进行第一阶段任务时，以序列总长度的 15%向下取整作为最大掩码数 m ，每个序列至少存在一个掩码。并在进行 MLM 任务时分以下两种情况：1) 患者存在心衰诊断编码时，假设包含心衰诊断数为 n ，则至少随机抽取相关心衰诊断 $1 \sim n$ 个，满足 $n \leq m$ ，并以 0.8 的概率替换为[MASK]，0.1 的概率保留，0.1 的概率替换为任意编码。剩余的 $m-n$ 个编码在非心衰编码位置继续进行补齐，2) 患者不存在心衰诊断时，按一般方式进行 MLM 训练。存在心衰诊断患者与非心衰诊断患者最终的掩码数均为 m 。第一阶段的输入构成有患者的既往诊断史编码、患者年龄、诊断优先级、就诊相对顺序。模型整体结构及预训练和微调差异见图 5.4。考虑到患者年龄、就诊相对顺序均能提供输入的位置信息以及时间信息，因此不再引入其余分隔标记类型。

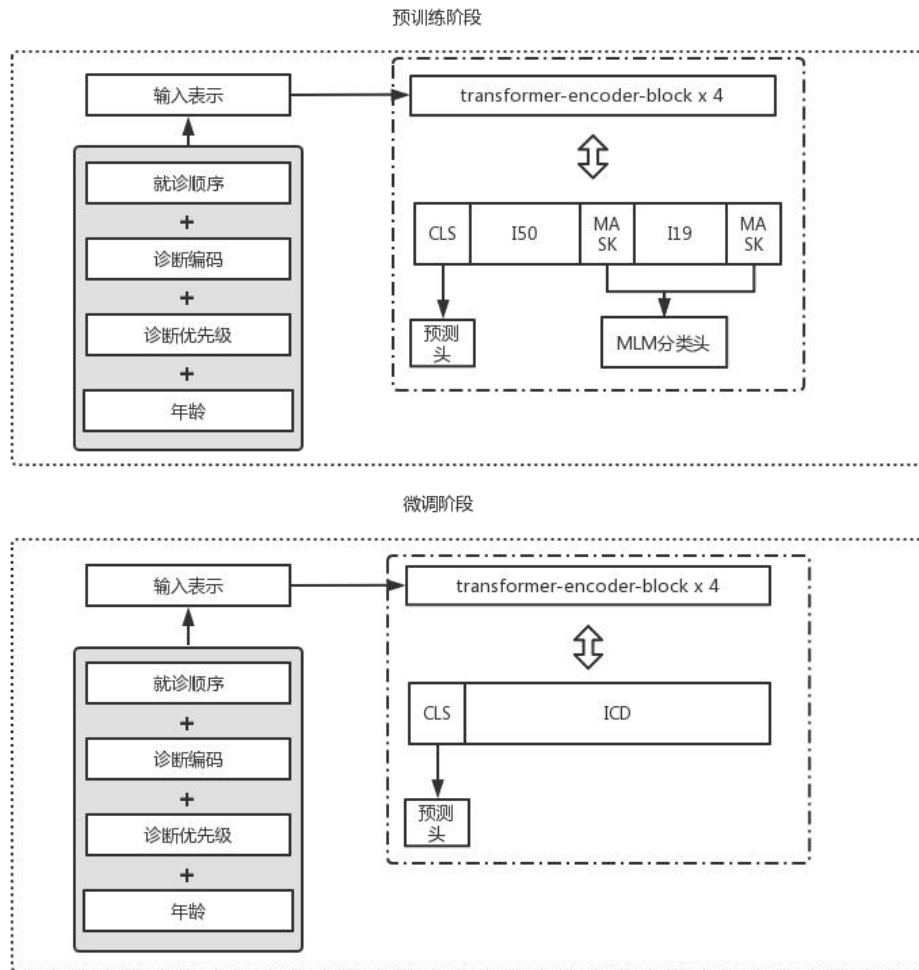


图 5.4 SBFHF 模型结构

Figure 5.4 Structure of SBFHF

在第二阶段任务时，发现心衰患者辨别任务相对 MLM 更加容易，因此在实际训练过程中通过增加调制参数 γ 减缓预测头及相关计算图路径上的参数更新梯度，防止模型过早出现过拟合现象。

在第三阶段任务时，考虑到预训练同样使用[CLS]进行心衰诊断识别任务，为区分微调阶段的死亡风险预测任务，将预测头的参数重新初始化并跟随网络训练，仅保留编码器部分权重。

训练过程中使用的优化器为 AdamW (Adaptive moment estimation)，余弦退火重启 (cosine annealing warm restarts) 作为学习率优化策略，预训练及微调阶段均使用早停技巧。预训练任务的损失函数由两部分构成，见下公式 (5-1~5-3)。第一部分是 MLM 的损失，由交叉熵计算得来。第二部分为心衰患者识别的分类损失，由公式 (4-3) 代入计算。

$$\text{Loss}_{\text{cls}_{\text{hf}}} = \sum_i^i \text{FocalLoss}(\text{hf}_i, p_2^i) \quad (5-1)$$

$$\text{Loss}_{\text{MLM}} = - \sum_{i=1}^n \sum_{j=1}^k y_{i,j} \log(p_{i,j}) \quad (5-2)$$

$$\text{Loss}_{\text{pretrain}} = \text{Loss}_{\text{MLM}} + \gamma \cdot \text{Loss}_{\text{cls}_{\text{hf}}} \quad (5-3)$$

算法 2 SBFHF 训练微调过程

输入: 预训练样本集 $S_1 = \{(\text{hf}_1, y^1, v^1_{\text{ehr_diag_icd}}, v^1_{\text{ehr_diag_order}}, v^1_{\text{ehr_hadm_order}}, v^1_{\text{ehr_age_order}}), \dots\}$, 其中 $v^1_{\text{ehr_diag_icd}}$ 为序列编码, $v^1_{\text{ehr_diag_order}}$ 为序列诊断顺序; $v^1_{\text{ehr_hadm_order}}$ 为序列就诊顺序; $v^1_{\text{ehr_age_order}}$ 为序列对应年龄编码; y^i 为患者是否心衰出院 1 年后死亡标签; hf_i 为患者是否存在心衰诊断; SBFHF 模型 f ; 最大预训练周期 T_{max} ; 早停数 patience ; 训练批大小 bs_size ; 训练集大小 N ; 历史损失记录 pre_loss ; 目标样本集 $S_2 = \{(\text{hf}_1, y^1, v^1_{\text{ehr_diag_icd}}, v^1_{\text{ehr_diag_order}}, v^1_{\text{ehr_hadm_order}}, v^1_{\text{ehr_age_order}}), \dots\}$, S_2 是 S_1 的真子集

输出: 患者的心衰出院后一年内死亡风险 $\text{output}_{\text{risk}}$

```

1: for 0  $\rightarrow$  4 do
2:   将预训练样本集  $S_1$  随机划分为训练集  $S'_{\text{train}}$  以及验证集合  $S'_{\text{eval}}$ , 置  $\text{pre\_loss} = 1e8$ 
3:   for 0  $\rightarrow$   $T_{\text{max}} - 1$  do
4:     for 0  $\rightarrow$   $N // \text{bs\_size}$  do
5:       从训练集合  $S'_{\text{train}}$  抽取  $\text{bs\_size}$  的样本  $S'$ , 并进行掩码任务处理
6:       计算  $S'$  的心衰诊断 loss 掩码任务 loss 使用 AdamW 进行梯度下降及优化
7:     end for
8:     计算验证集合上  $S'_{\text{eval}}$  的损失  $\text{eval\_loss}$ 
9:     if  $\text{eval\_loss} \geq \text{pre\_loss}$  then  $\text{cur\_patience} += 1$ 
10:      else  $\text{cur\_patience} = 0$ 
11:    end if
12:     $\text{pre\_loss} = \text{epoch\_loss}$ 
13:    if  $\text{patience} == \text{cur\_patience}$  then 停止预训练
14:    else 继续预训练
15:  end if
16: end for
17: 保留  $f$  的特征提取层参数, 重新初始化  $f$  的分类层参数, 在  $S_2$  数据集上进行微调时
    步骤见上 3 ~ 17, loss 计算仅使用  $y^i$  标签, 不再进行掩码预测
18: end for

```

5.3 实验结果

预训练时训练集和验证集比例为 0.8:0.2, 微调阶段的训练、验证、测试比例为

0.7:0.1:0.2。在预训练阶段及微调阶段的数据可能有一定重合度，后续结果均基于 5 折交叉方式，尽量避免数据重叠。本节将从心衰患者的死亡预测任务上评估预训练范式的有效性。模型参数配置上，注意力头数为 6，模型输入维度为 120，前馈神经网络维度为 64，堆叠 4 层 Transformer 编码器。为验证输入标记类型对于模型性能的影响，将不同的 7 种组合方式及对应模型表现性能进行统计（表 5.4）。表中数据均在使用临床诊断输入标记基础上，评估年龄（A）、就诊顺序（H）、诊断优先级（P）三种额外输入标记对最终的影响。如 AHP 代表使用三种输入标记。

表 5.4 不同类型模型性能比对

Table 5.4 Performance comparison of different types of models

Type	AUROC-pretrained	AUPRC-pretrained	AUROC-wto	AUPRC-wto	AUROC-MLM	AUPRC-MLM
AH	0.688	0.267	0.647	0.236	0.680	0.264
AP	0.689	0.268	0.656	0.252	0.676	0.250
AHP	0.708	0.287	0.636	0.232	0.698	0.283
PH	0.719	0.284	0.636	0.232	0.688	0.255
H	0.725	0.305	0.672	0.251	0.705	0.285
P	0.720	0.304	0.694	0.264	0.708	0.290
A	0.705	0.278	0.657	0.244	0.687	0.250

Type 为纳入标记类型组合缩写；pretrained 代表该列评估数据经过预训练；wto 代表未经过预训练的模型；MLM 代表未使用代理任务二的对比模型。

从表 5.4 的左侧 5 列可知，经过预训练的模型相比未预训练都有显著的性能提升。但是从不同标记引入类型来看，差别较为明显。分析是由于每一种输入类型标记的值域范围不同。按照 5.1 节预定义范围，A 的范围在（18~89，91）之间，P 的范围在（1~39），H（1~8）的范围。可知 H 标记的学习所需样本量是相对较少的，因此对于输入仅包含 H 的模型而言，在训练样本量相同条件下，其性能是相对高的，并且按照样本量需求的递减排列后的 A、P、H 三者对应的模型性能呈现了递增的趋势，猜测是目前样本量下模型的训练可能是不够充分的。然而，同时引入三种输入标记后，模型性能处于中等水平，这种现象发生的原因可能是当三种类型输入与诊断信息共同表示输入时，其表达的语义信息是最符合诊断序列中应涵盖的时间、顺序、临床诊断意义。另为证明代理任务二对于整体模型表现的影响，对比仅使用 MLM 任务及联合代理任务二下的不同模型表现性能。根据表 5.4 的右侧两列及左侧两列对比发现，未使用代理任务二（右侧两列）相比下都略低于联合训练时（左侧两列）的表现性能，证明了代理任务二增强了心衰患者死亡风险的预测准确度。

但从整体的实验结果来看，模型的性能是较不理想的。为发掘不理想结果的原因，

另从预训练及目标数据的各项分布上进行分析（图 5.5）。预训练的成功可看作是基于一大规模数据训练的模型将其可感知的数据分布尽可能扩大，使得下游任务预测时所见数据尽可能在模型的可理解范畴内。据图 5.5 可知，在目标患者中所使用的输入标记类型及序列长度的整体分布都与预训练数据的整体分布存在差异，目标诊断的数据将更为复杂，这可能也是导致结果不理想的原因之一。

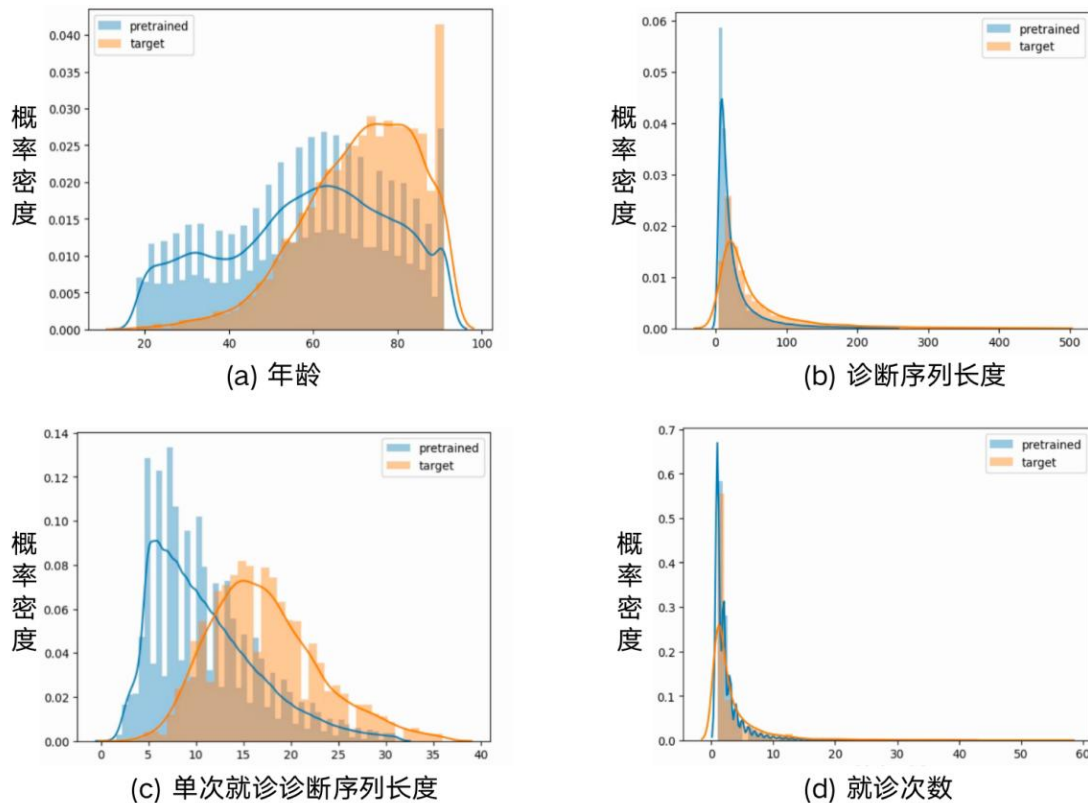


图 5.5 预训练数据与目标研究数据间的差异

Figure 5.5 Discrepancies between pre-training data and target study data

为探究预训练数据集大小对于模型最终效果的影响，将 eICU^[35]中的全体诊断数据按照 5.1 节的筛选条件进行提取，共获得 59,057 名患者的 2,276,209 条诊断数据。其中存在心衰诊断患者 13,536 名，IR 为 3.36，但提取的诊断数据记录方式与 MIMIC-IV 不同（表 5.5）。其中住院标示号唯一确定了患者 ICU 入院的信息；是否出院诊断表示当前诊断信息是否是 ICU 出院前的末次诊断；诊断标示号是诊断表中的主键；相对 ICU 入院表示该项诊断是相对 ICU 入院的多少分钟差值，该值可能存在负值，表示进入 ICU 前的诊断信息；诊断描述则是诊断编码的更详细文字描述，以“|”间隔不同描述，如肺 | 气道疾病 | 慢性阻塞性肺部疾病；诊断编码中包含 ICD-9 及 ICD-10；诊断优先级包括“Major”、“Primary”、“Other”三种类型，前两项都代表主诊断。

对于 SBFHF 的输入标记，将相对 ICU 入院时间作为诊断序列顺序（P）构造基础，值越小时表明产生诊断越早。就诊顺序（H）可通过同一患者不同住院标示号间的大小

进行排序获取。年龄 (A) 可直接在 eICU 的患者表中获取。

表 5.5 eICU 诊断表字段介绍

Table 5.5 Description of diagnosis fields in eICU

名称	含义	名称	含义
patientUnitStayID	住院标示号	diagnosisString	诊断描述
activeUponDischarge	是否出院诊断	ICD9Code	诊断编码
diagnosisID	诊断标示号	diagnosisPriority	诊断优先级
diagnosisOffset	相对 ICU 入院时间 (分)	-	-

合并 eICU 数据后, 训练集及验证集分别含 174,542, 43,635 名患者, 并按相同方式进行模型预训练, 最终 5 折验证对比性能见图 5.6。图中 ROC 为 AUROC 大小。

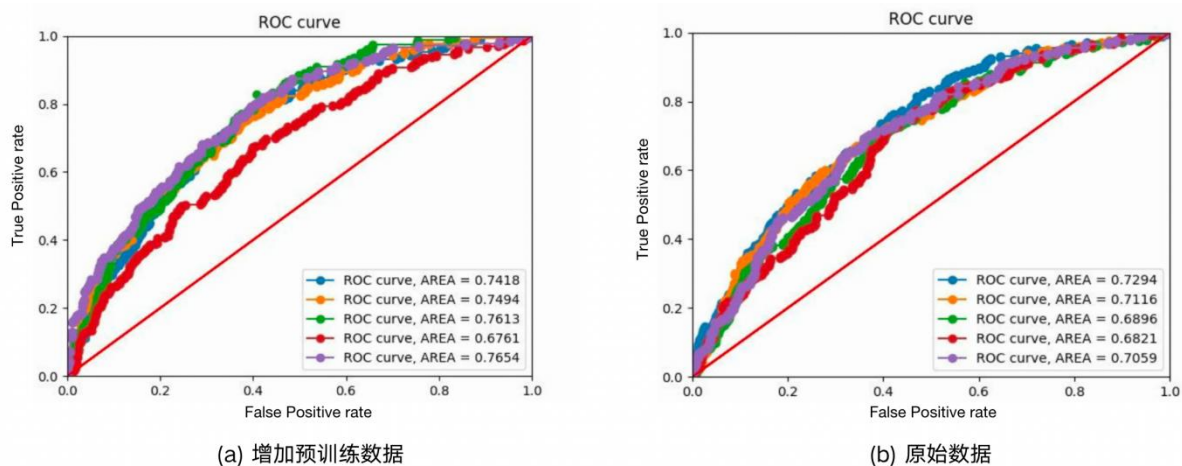


图 5.6 不同预训练数据规模下的 SBFHF 性能

Figure 5.6 SBFHF performance at different pretraining data scales

据图 5.6 可知, 增加预训练数据规模后的心衰患者死亡风险预测任务中的 4 次性能均大于原始数据规模下的所有模型表现性能, 并且提升性能是显著的, 其中一次性能略低可能受到了测试样本难度的影响, 整体来看模型仍可通过增加预训练数据量的形式逐步提升预测准确度, 也证明了模型性能不理想的原因其一。

另外, 为比较不同模型下的临床评估效益, 将不同输入标记类型及增加预训练数据规模后的模型及对应的临床决策分析曲线统一绘制 (图 5.7)。其中 AHP_eICU 代表增加 eICU 诊断数据后的对应预训练模型, 其余标记缩写与前文一致。

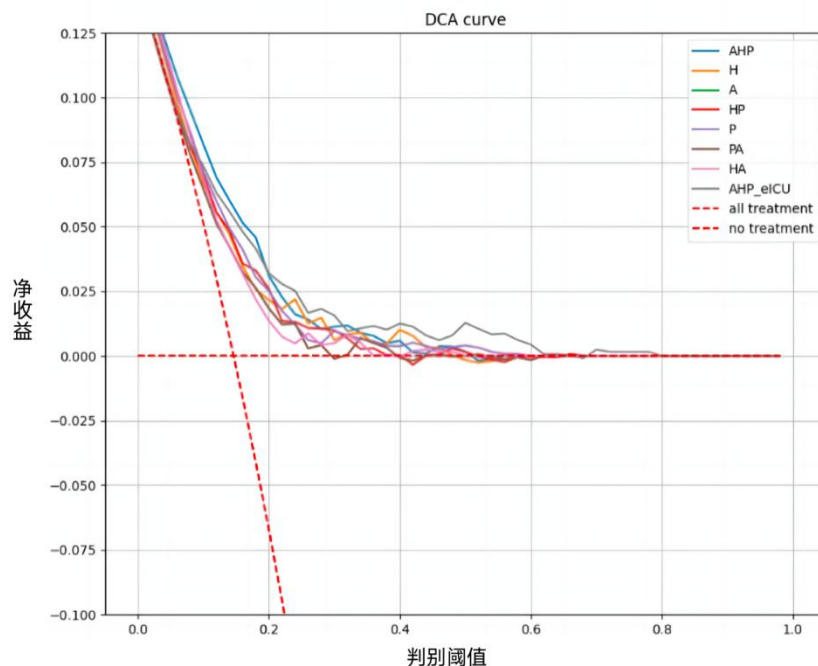


图 5.7 不同输入的 DCA 曲线

Figure 5.7 DCA curves with different inputs

从各曲线的净收益来看，引入三种标记的模型（AHP）及增加预训练数据规模（AHP_eICU）的表现性能是最优的，且 AHP_eICU 在相对较高判别阈值下的表现也是一直优于整体治疗的状况，进一步证明了数据集规模的增大将持续提升 SBFHF 的临床应用表现能力。

为比较 SBFHF 对于患者死亡风险的预测输出合理性及应用价值，将不同临床结局患者分类为生存（alive）、30 天内死亡（ ≤ 30 ）、360 天内死亡（ ≤ 360 ）、360 天以外死亡的四种，并将模型的预测死亡风险通过箱形图方式对比展示（图 5.8）。据图 5.8 显示，虽然模型对于生存患者的死亡风险还存在异常高风险输出，但从四种患者群体的预测均值而言，SBFHF 的输出是合理的。在死亡患者中，30 天内死亡患者的平均输出死亡风险是大于 360 天以内及以外的患者，且上下四分位数的中间区域数据也是随着死亡时间与心衰入院的时间间隔增大而减小，其死亡风险概率的平均值也同时是减小的。理论上，患者的真实死亡时间越接近于心衰入院时间时，其死亡风险也应该越大，SBFHF 的输出也符合该趋势。

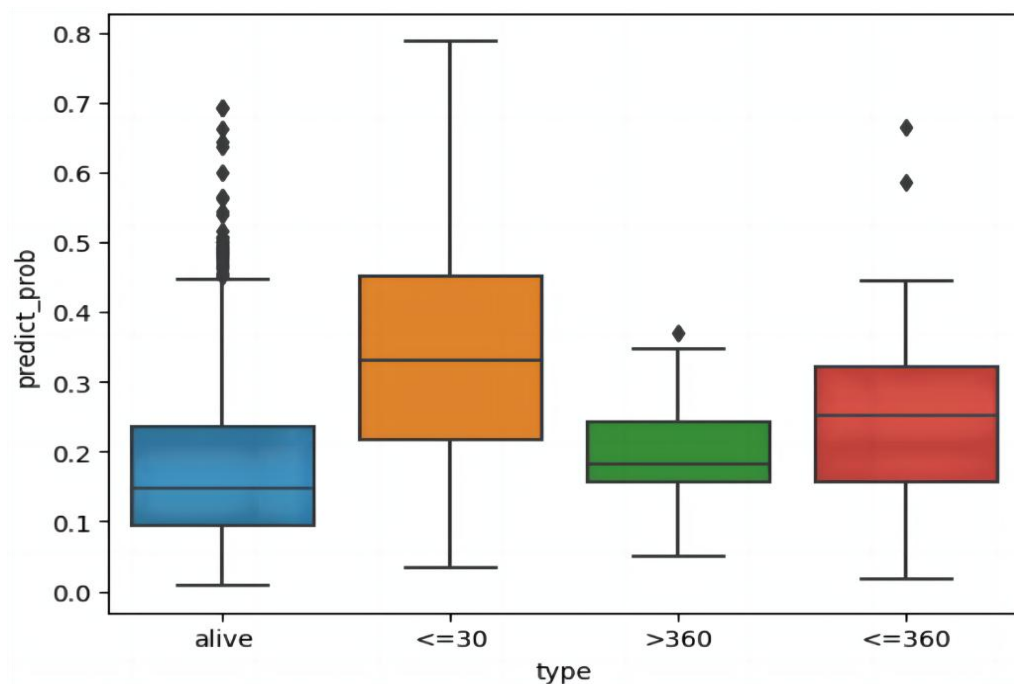


图 5.8 不同类型患者的死亡风险输出

Figure 5.8 Output of mortality risk for different types of patients

5.4 解释性分析

Transformer 结构中的多头注意力模块是自带解释性意义的。通过可视化患者死亡风险推理过程中的网络注意力分布，可知与目标结果高相关的诊断。预训练阶段[CLS]由于代理任务 2 的因素，已具备了序列的整体感知能力，因此将[CLS]标记与其余诊断的注意力作为相关性分析依据（图 5.9），已分析死亡风险预测与其余诊断的相关性。注意力绘图基于 bertviz^[59]工具。



图 5.9 [CLS]的注意力分布

Figure 5.9 Attention distribution of [CLS]

图 5.9 中，左上角六种方框颜色对应不同的注意力头，颜色深浅代表该头下注意力的大小。在网络浅层时[CLS]与心衰（I50）等具有相对较大关联，在网络的第六层时，输入的注意力间关系更加具体化。[CLS]与心衰（I50），体液、电解质、酸碱平衡紊乱（E87），其他肝脏疾病（K76）具有更大的关系。由于患者存在多次就诊，因此在序列中可能出现相同诊断多次，在该患者就诊的诊断序列中，出现了三次心衰诊断，而后两次的诊断对于患者死亡风险提供了更大的影响。

除此之外，根据网络整体的注意力分布，也能发现一些强相关的诊断信息（图 5.10）。图 5.10 中线条颜色深浅代表了注意力的大小，可发现心衰与未指明病原体的肺炎（J16）存在一定的关系。与邬真力报告的心衰患者与肺炎具有很大关联一致^[61]。

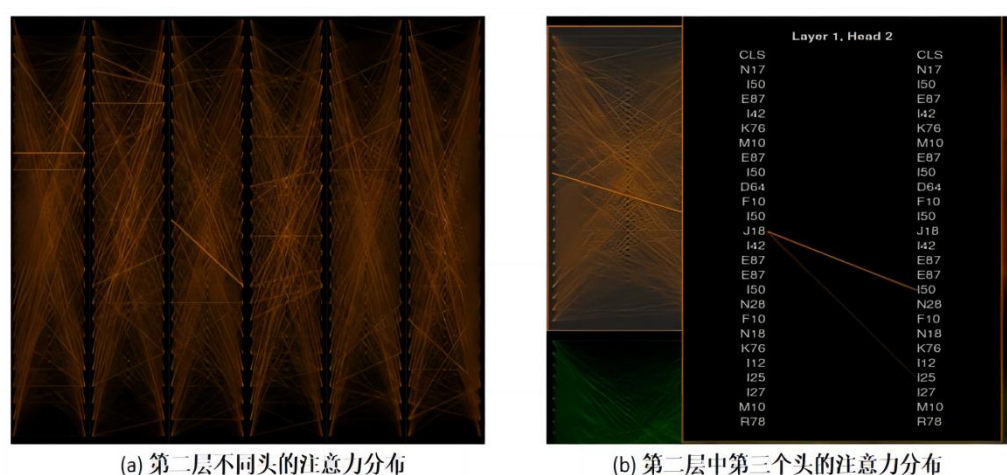


图 5.10 SBFHF 的第二层的整体注意力及局部注意力分布

Figure 5.10 Global attention and partial attention distribution in the second layer of SBFHF

5.5 本章小结

本章节提出面向心衰患者的预训练模型，并使用了两种预训练代理任务，分别是改进的 MLM 任务以及心衰诊断患者的辨别任务。预训练代理任务的设计使得 SBFHF 算法无需额外人工标注信息，降低了临床算法的使用门槛。从临床分析层面来看，可在患者刚入院时结合既往诊断给予风险评估，更加便捷。另外，通过实验结果证明了代理任务能有效指导模型提取心衰患者有用的表示特征，增强患者死亡风险的预测能力。同时 SBFHF 也能发掘诊断信息间一定关联的特点，如心衰与肺炎间的关联，

就模型性能而言，还存在较大提升空间。为探究数据量与模型表现力的关系，额外引入 eICU 的诊断数据并通过实验证明了 SBFHF 的可提升潜力。同时经过数据分布分析也发现了模型性能目前的瓶颈所在。

第6章 总结和展望

针对心衰患者的研究层出不穷,有效利用临床数据,基于回顾性的临床数据分析,为患者更精确的预后评估及治疗提供了契机。本研究提出的算法分别基于短时间内的ICU多源数据以及基于长周期的诊断数据两个角度来预测心衰患者的未来死亡风险,有助于临床医师评估心衰患者的预后状态。

在第四章所提出的TFA-GRU网络中,通过挖掘患者的临床干预模式变化及生理指标以判断患者潜在死亡风险。在基于ICU内24小时监护数据下,算法达到了较好的预测性能。但在建模过程中,仍然存在一些不足,如未引入用药信息。用药信息表构成较为复杂,包括营养剂、抗生素、血液制品、用药速率、用药剂量以及药物配比浓度等。引入该表需较专业的临床用药知识,且用药速率及用药剂量等指标的计算型表示相对其余输入特征来说较为不同,因此该表暂未纳入在输入信息源中。但从结果分析来看,TFA-GRU网络与临床研究结果相符,也能挖掘出一些与心衰患者死亡风险相关的因素,如动脉血二氧化碳分压的大小。总结来看,无论是从临床风险指标发掘、模型输出的合理性分析而言,TFA-GRU网络是一次在临床中有效的深度学习研究应用。但是基于电子病历的时序数据研究,是十分依赖各项临床指标完整性的。在MIMIC-IV中,数据缺失状况是较为严重的。另外,检验项目定义组织不清晰,如排出表内的导管指标表意不明,需花费更多时间组织整理数据结构,预将其作为短时间风险研究的数据集标杆还有待探究。HiRID^[62]数据集是每2分钟记录的高分辨率ICU临床数据,涵盖了55,602名患者的7千余项临床指标,但由于并未获取其数据使用权限,本文未能使用其数据进行研究,但该数据集具有极高的研究价值。相比其余开源临床数据而言,其高分辨率的数据或能给TFA-GRU网络提供更精确的数据支持,以输出更精确的患者预后评估。

第五章提出的SBFHF模型基于历史诊断、患者年龄、诊断优先级信息构建算法。相比传统医学统计量表使用了更少的临床指标,但与目前多数深度学习方法相比,预测准确度还有很大提升空间。基于Transformer算法的模型或许需上百万的语料数据,才能发挥更佳的性能。SBFHF在引入额外约6万名患者的诊断数据后,性能约提升了3%。然而目前基于近20万患者数据下,就已经达到了不错的预测水平,未来若能获取到更大规模、更高质量的数据,SBFHF模型架构能实现更好的性能。从实验结果来看,SBFHF的预训练范式能给心衰患者死亡风险预测带来性能的提升,也能将其推广到其他疾病患者的死亡风险预测中。第五章研究仅代表基于电子病历数据的临床预训练模型的一种尝试,结果也证明了迁移学习的范式是可以应用在电子病历中的,并且能在仅基于较少诊断信息的前提下,有效评估患者死亡风险。

除了模型以外，MIMIC-IV 数据库的版本也在不断更新，后续版本中模块进行了简化，部分表结构进行了优化，未来研究者可考虑相较 v1.0 更新的版本。不但有单中心的患者数据，还将纳入多中心的州随访数据，这也导致 v2.0 版本额外补充 15,621 名 ICU 的死亡患者。补充的死亡患者登记或能解释部分生存患者在模型预测下出现了较高的死亡风险原因。

本文研究均聚焦于 ICU 住院患者，主要由于 ICU 患者的监护设备齐全，数据丰富，且都是病情危重的患者。后续实验可考虑非 ICU 的患者数据，利用尽可能少的临床指标，达到更高的模型性能，并在外部数据库验证模型有效性。

总结来看，深度学习方法能灵活结合多源数据的分析，充分挖掘潜在的风险模式，有助于临床知识的发现。这是传统机器学习方法以及医学统计方法很难做到的。本研究通过可视化模型输出的方法，证明了深度模型在临床应用的潜力。

参考文献

- [1] Li M, Gao X, Wang H, et al. Phosphoglycerate mutase 2 is elevated in serum of patients with heart failure and correlates with the disease severity and patient's prognosis[J]. Open Medicine, 2021, 16(1): 1134-1142.
- [2] Zaharova S, Litwack K, Gopalakrishnan S, et al. Self-management in heart failure: the importance of self-regulation but not complexity of condition[J]. Western Journal of Nursing Research, 2022, 44(4): 375-382.
- [3] Bragazzi N L, Zhong W, Shu J, et al. Burden of heart failure and underlying causes in 195 countries and territories from 1990 to 2017[J]. European Journal of Preventive Cardiology, 2021, 28(15): 1682-1690.
- [4] Savarese G, Lund L H. Global public health burden of heart failure[J]. Cardiac failure review, 2017, 3(1): 7.
- [5] Groenewegen A, Rutten F H, Mosterd A, et al. Epidemiology of heart failure[J]. European journal of heart failure, 2020, 22(8): 1342-1356.
- [6] 中国心血管健康与疾病报告编写组. 中国心血管健康与疾病报告 2019 概要[J]. 中国循环杂志, 2020, 35(9): 833-854.
- [7] Park L G, Dracup K, Whooley M A, et al. Sedentary lifestyle associated with mortality in rural patients with heart failure[J]. European Journal of Cardiovascular Nursing, 2019, 18(4): 318-324.
- [8] Bozkurt B, Coats A J, Tsutsui H, et al. Universal definition and classification of heart failure: a report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure[J]. Eur J Heart Fail, 2021, 23: 352-380.
- [9] 张永珍, 范媛媛. 心力衰竭通用定义和分类解读[J]. 中国心血管杂志, 2021, 26(5): 409-412.
- [10] Rahimi K, Bennett D, Conrad N, et al. Risk prediction in patients with heart failure: a systematic review and analysis[J]. JACC Heart Fail. 2014, 2(5): 440-446.
- [11] 中国医疗保健国际交流促进会急诊医学分会, 中华医学会急诊医学分会, 中国医师协会急诊医师分会, 等. 急性心力衰竭中国急诊管理指南[J]. 临床急诊杂志, 2022, 23(8): 29.
- [12] Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study[J]. bmj, 2017, 357.
- [13] Kannel W B, Dawber T R, Kagan A, et al. Factors of risk in the development of coronary heart disease—six-year follow-up experience: the Framingham Study[J]. Annals of internal medicine, 1961, 55(1): 33-50.
- [14] Woodward M, Brindle P, Tunstall-Pedoe H. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC)[J]. Heart, 2007, 93(2): 172-176.
- [15] Levy W C, Mozaffarian D, Linker D T, et al. The Seattle Heart Failure Model: prediction of survival in heart failure[J]. Circulation, 2006, 113(11): 1424-1433.
- [16] Vazquez R, Bayes-Genis A, Cygankiewicz I, et al. The MUSIC Risk score: a simple method

- for predicting mortality in ambulatory patients with chronic heart failure[J]. *European heart journal*, 2009, 30(9): 1088-1096.
- [17] Miró Ò, Rossello X, Gil V, et al. Predicting 30-day mortality for patients with acute heart failure in the emergency department: a cohort study[J]. *Annals of internal medicine*, 2017, 167(10): 698-705.
- [18] Miró Ò, Rossello X, Platz E, et al. Risk stratification scores for patients with acute heart failure in the Emergency Department: a systematic review[J]. *European Heart Journal: Acute Cardiovascular Care*, 2020, 9(5): 375-398.
- [19] Lu S, Chen R, Wei W, et al. Understanding heart failure patients EHR clinical features via SHAP interpretation of tree-based machine learning model predictions[C]//AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2021, 2021: 813.
- [20] Li J, Liu S, Hu Y, et al. Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: retrospective cohort study[J]. *Journal of Medical Internet Research*, 2022, 24(8): e38082.
- [21] Luo C, Zhu Y, Zhu Z, et al. A machine learning-based risk stratification tool for in-hospital mortality of intensive care unit patients with heart failure[J]. *Journal of Translational Medicine*, 2022, 20(1): 136.
- [22] Zaman S M M, Qureshi W M, Raihan M M S, et al. Survival Prediction of Heart Failure Patients using Stacked Ensemble Machine Learning Algorithm[C]//2021 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE). IEEE, 2021: 117-120.
- [23] Takahama H, Nishimura K, Ahsan B, et al. Clinical application of artificial intelligence algorithm for prediction of one-year mortality in heart failure patients[J]. *Heart and Vessels*, 2023: 1-8.
- [24] 姚丽娟, 李冬冬, 王喆. 基于 Relief 特征选择的心衰死亡率预测[J]. *计算机工程与应用*, 2018, 54(23):125-130.
- [25] Tunc Z, Balıkcı C İ, Güldoğan E, et al. Assessment of associative classification approach for predicting mortality by heart failure[J]. *The Journal of Cognitive Systems*, 2020, 5(2): 41-45.
- [26] PARK J, HWANG I, YOON Y, et al. Predicting Long-Term Mortality in Patients With Acute Heart Failure by Using Machine Learning[J]. *Journal of Cardiac Failure*, 2022, 28(7): 1078-1087.
- [27] Choi E, Bahadori M T, Schuetz A, et al. Doctor ai: Predicting clinical events via recurrent neural networks[C]//Machine learning for healthcare conference. PMLR, 2016: 301-318.
- [28] Che Z, Purushotham S, Cho K, et al. Recurrent Neural Networks for Multivariate Time Series with Missing Values[J]. *Scientific Reports*, 2018, 8(1): 6085.
- [29] 刘卢琛, 沈剑豪, 张铭, 等. 基于深度学习的异构时序事件患者数据表示学习框架[J]. *大数据*, 2019, 5(1): 14.
- [30] Nguyen P, Tran T, Wickramasinghe N, et al. Deepr: A Convolutional Net for Medical Records[J]. *IEEE J Biomed Health Inform*, 2017, 21(1): 22-30.
- [31] Li D, Fu J, Zhao J, et al. A deep learning system for heart failure mortality prediction[J]. *PLoS one*, 2023, 18(2): e0276835.
- [32] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. *Advances in neural information processing systems*, 2017, 30.
- [33] Rasmy L, Xiang Y, Xie Z, et al. Med-BERT: pretrained contextualized embeddings on large-s

- cale structured electronic health records for disease prediction[J]. NPJ digital medicine, 2021, 4(1): 86.
- [34] Johnson A E W, Pollard T J, Shen L, et al. MIMIC-III, a freely accessible critical care data base[J]. Scientific data, 2016, 3(1): 1-9.
- [35] Pollard T J, Johnson A E W, Raffa J D, et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research[J]. Scientific data, 2018, 5(1): 1-13.
- [36] Herrett E, Gallagher A M, Bhaskaran K, et al. Data resource profile: clinical practice research datalink (CPRD)[J]. International journal of epidemiology, 2015, 44(3): 827-836.
- [37] Chisholm J. The Read clinical classification[J]. BMJ: British Medical Journal, 1990, 300(6732): 1092.
- [38] Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age[J]. PLoS medicine, 2015, 12(3): e1001779.
- [39] BioBank. Data-Field 131355[DB/OL].(2019.7)[2023.3].<https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=131355>
- [40] Ahmad T, Munir A, Bhatti S H, et al. Survival analysis of heart failure patients: A case study[J]. PloS one, 2017, 12(7): e0181001.
- [41] Johnson A, Bulgarelli L, Pollard T, et al. Mimic-iv[J]. PhysioNet. Available online at: <https://physionet.org/content/mimiciv/1.0/>(accessed August 23, 2021), 2020.
- [42] Fonarow G C. The Acute Decompensated Heart Failure National Registry (ADHERE): opportunities to improve care of patients hospitalized with acute decompensated heart failure[J]. Reviews in cardiovascular medicine, 2003, 4: S21-30.
- [43] Johnson A E W, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset[J]. Scientific data, 2023, 10(1): 1.
- [44] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [45] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[J]. Advances in neural information processing systems, 2017, 30.
- [46] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
- [47] Vickers A J, Elkin E B. Decision curve analysis: a novel method for evaluating prediction models[J]. Medical Decision Making, 2006, 26(6): 565-574.
- [48] Tomašev N, Glorot X, Rae J W, et al. A clinically applicable approach to continuous prediction of future acute kidney injury[J]. Nature, 2019, 572(7767): 116-119.
- [49] Adler E D, Voors A A, Klein L, et al. Improving risk prediction in heart failure using machine learning[J]. European journal of heart failure, 2020, 22(1): 139-147.
- [50] Wu Z, Xiong Y, Yu S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3733-3742.
- [51] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
- [52] 刘兆龙, 叶玲玲. 慢性心力衰竭患者血尿素氮升高 198 例临床分析[J]. 中国基层医药, 2014, (10): 1576-1577.

- [53] 安嫵, 侯丹华, 范秀珍. 血清氯离子浓度与心力衰竭患者死亡风险的 Meta 分析[J]. 中国老年学杂志, 2021, 41(20): 4339-4344.
- [54] 华琦, 李东宝, 万云高, 等. 老年急性心肌梗死患者白细胞计数与心脏功能及在院死亡率的关系[J]. 中华老年多器官疾病杂志, 2007, 6(5): 305-307.
- [55] Pang C, Jiang X, Kalluri K S, et al. CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks[C]//Machine Learning for Health. PMLR, 2021: 239-260.
- [56] Li Y, Mamouei M, Salimi-Khorshidi G, et al. Hi-BEHRT: Hierarchical Transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records[J]. IEEE Journal of Biomedical and Health Informatics, 2022.
- [57] Grill J B, Strub F, Althé F, et al. Bootstrap your own latent-a new approach to self-supervised learning[J]. Advances in neural information processing systems, 2020, 33: 21271-21284.
- [58] Li Y, Rao S, Solares J R A, et al. BEHRT: transformer for electronic health records[J]. Scientific reports, 2020, 10(1): 1-12.
- [59] Rao S, Li Y, Ramakrishnan R, et al. An Explainable Transformer-Based Deep Learning Model for the Prediction of Incident Heart Failure[J]. IEEE J Biomed Health Inform, 2022, 26(7):3362-3372.
- [60] Vig J . A Multiscale Visualization of Attention in the Transformer Model[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2019,37-42.
- [61] 邬真力. 近一半的心衰患者合并肺炎, 如何抗感染?[EB/OL]. [2018-3-26]https://news.medlive.cn/heart/info-progress/show-140619_129.html
- [62] Hyland S L, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning[J]. Nature medicine, 2020, 26(3): 364-373.

附录 A

附录 A-1 心衰诊断定义表

ICD 编码	ICD 版本	ICD 编码	ICD 版本
39891	9	I50	10
40201	9	I502	10
40211	9	I5020	10
40291	9	I5021	10
40401	9	I5022	10
40403	9	I5023	10
40411	9	I503	10
40413	9	I5030	10
40491	9	I5031	10
40493	9	I5032	10
4280	9	I5033	10
4281	9	I504	10
42820	9	I5040	10
42821	9	I5041	10
42822	9	I5042	10
42823	9	I5043	10
42830	9	I508	10
42831	9	I5081	10
42832	9	I50810	10
42833	9	I50811	10
42840	9	I50812	10
42841	9	I50813	10
42842	9	I50814	10
42843	9	I5082	10
4289	9	I5083	10
I509	10	I5084	10
I0981	10	I5089	10

附录 A-1 续

I110	10	I9713	10
I130	10	I97130	10
I132	10	I97131	10

附录 A-2 纳入特征及数据库记录特性

Itemid	Label_name	linksto	Param_type	Category
220048	Heart Rhythm	chartevents	Text	Routine Vital Signs
223753	Riker-SAS Scale	chartevents	Text	Pain/Sedation
227341	History of falling (within 3 mnths)	chartevents	Text	Restraint/Support Systems
228096	Richmond-RAS Scale	chartevents	Text	Pain/Sedation
228299	Goal Richmond-RAS Scale	chartevents	Text	Pain/Sedation
228332	Delirium assessment	chartevents	Text	Pain/Sedation
220045	Heart Rate	chartevents	Numeric	Routine Vital Signs
220050	Arterial Blood Pressure systolic	chartevents	Numeric	Routine Vital Signs
220051	Arterial Blood Pressure diastolic	chartevents	Numeric	Routine Vital Signs
220052	Arterial Blood Pressure mean	chartevents	Numeric	Routine Vital Signs
220059	Pulmonary Artery Pressure systolic	chartevents	Numeric	Hemodynamics
220060	Pulmonary Artery Pressure diastolic	chartevents	Numeric	Hemodynamics
220061	Pulmonary Artery Pressure mean	chartevents	Numeric	Hemodynamics
220074	Central Venous Pressure	chartevents	Numeric	Hemodynamics
220210	Respiratory Rate	chartevents	Numeric	Respiratory
220224	Arterial O2 pressure	chartevents	Numeric	Labs

附录 A-2 续

220228	Hemoglobin	chartevents	Numeric	Labs
220235	Arterial CO2 Pressure	chartevents	Numeric	Labs
220277	O2 saturation pulseoxymetry	chartevents	Numeric	Respiratory
220339	PEEP set	chartevents	Numeric	Respiratory
220545	Hematocrit (serum)	chartevents	Numeric	Labs
220546	WBC	chartevents	Numeric	Labs
220602	Chloride (serum)	chartevents	Numeric	Labs
220615	Creatinine (serum)	chartevents	Numeric	Labs
220621	Glucose (serum)	chartevents	Numeric	Labs
220635	Magnesium	chartevents	Numeric	Labs
220645	Sodium (serum)	chartevents	Numeric	Labs
223762	Temperature Celsius	chartevents	Numeric	Routine Vital Signs
223772	SvO2	chartevents	Numeric	Hemodynamics
223830	PH (Arterial)	chartevents	Numeric	Labs
224639	Daily Weight	chartevents	Numeric	General
224828	Arterial Base Excess	chartevents	Numeric	Labs
224842	Cardiac Output (CCO)	chartevents	Numeric	Hemodynamics
225624	BUN	chartevents	Numeric	Labs
225625	Calcium non-ionized	chartevents	Numeric	Labs
225667	Ionized Calcium	chartevents	Numeric	Labs
225668	Lactic Acid	chartevents	Numeric	Labs

附录 A-2 续

225677	Phosphorous	chartevents	Numeric	Labs
225698	TCO2 (calc) Arterial	chartevents	Numeric	Labs
226730	Height (cm)	chartevents	Numeric	General
227073	Anion gap	chartevents	Numeric	Labs
227442	Potassium (serum)	chartevents	Numeric with tag	Labs
227443	HCO3 (serum)	chartevents	Numeric with tag	Labs
227446	Brain Natriuretic Peptide (BNP)	chartevents	Numeric with tag	Labs
227457	Platelet Count	chartevents	Numeric with tag	Labs
227465	Prothrombin time	chartevents	Numeric with tag	Labs
227466	PTT	chartevents	Numeric with tag	Labs
227467	INR	chartevents	Numeric with tag	Labs
227471	Specific Gravity (urine)	chartevents	Numeric with tag	Labs
225124	Unintentional weight loss >10 lbs.	chartevents	Checkbox	Adm History/FHPA
225401	Blood Cultured	procedureevents	Processes	6-Cultures
225464	Cardioversion/Defibrillation	procedureevents	Processes	3-Significant Events
225465	Chest Pain	procedureevents	Processes	3-Significant Events
225466	Cardiac Arrest	procedureevents	Processes	3-Significant Events
225475	Respiratory Arrest	procedureevents	Processes	3-Significant Events
225479	Thoracentesis	procedureevents	Processes	4-Procedures
229366	Epoprostenol (Inhaled)	procedureevents	Processes	Medications

附录 A-2 续

226627	OR Urine	outputevents	Numeric	Output
226559	Foley	outputevents	Numeric	Output
220179	Non Invasive Blood Pressure systolic	chartevents	Numeric	Routine Vital Signs
220180	Non Invasive Blood Pressure diastolic	chartevents	Numeric	Routine Vital Signs
220181	Non Invasive Blood Pressure mean	chartevents	Numeric	Routine Vital Signs
223761	Temperature Fahrenheit	chartevents	Numeric	Routine Vital Signs
224690	Respiratory Rate (Total)	chartevents	Numeric	Respiratory
225312	ART BP Mean	chartevents	Numeric	Routine Vital Signs
225664	Glucose finger stick (range 70-100)	chartevents	Numeric	Labs
226512	Admission Weight (Kg)	chartevents	Numeric	General
226531	Admission Weight (lbs.)	chartevents	Numeric	General
226534	Sodium (whole blood)	chartevents	Numeric	Labs
226537	Glucose (whole blood)	chartevents	Numeric	Labs
226707	Height	chartevents	Numeric	General
226846	Feeding Weight	chartevents	Numeric	General
227464	Potassium (whole blood)	chartevents	Numeric with tag	Labs
229761	Creatinine (whole blood)	chartevents	Numeric	Labs