# 一种用于心衰患者死亡率预测的数据多重插补方法*

张雅楠[1], 张琳琳[1,2], 郭渊博[1], 毕雪华[3], 赵 楷[4†]

(1. 新疆大学 软件学院,新疆 乌鲁木齐 830091; 2. 新疆大学 网络与信息技术中心,新疆 乌鲁木齐 830046;
3. 新疆医科大学 医学工程与技术学院,新疆 乌鲁木齐 830054; 4. 新疆大学 计算机科学与技术学院,新疆 乌鲁木齐 830017)

**摘 要**:针对真实数据采集机制不完善致使数据缺失、现有方法对临床特征表示不足导致模型性能受限问题,本文提出一种用于心衰患者死亡率预测的数据多重插补方法(Self-attention and Generative adversarial network based Mortality Prediction, SGMP). 首先,针对临床特征在变分自编码器(Variational Autoencoder, VAE)的潜在空间中结合自注意力机制动态融合多组候选估计值,并结合生成对抗网络(Generative Adversarial Network, GAN)的对抗训练策略优化表征学习能力. 然后,根据掩码矩阵有效获取候选估计结果,实现缺失数据多重插补. 最后,采用合成少数类过采样技术(Synthetic Minority Over-sampling Technique, SMOTE)进行数据增强,使用多层感知机(Multilayer Perceptron, MLP)实现死亡率预测. 基于新疆某三甲医院心衰患者数据进行验证,结果表明:死亡率预测任务中,相比其他模型,SGMP在多个指标上有明显提升,受试者工作特征曲线下面积达到0.902.

**关键词**:死亡率预测;多重插补(MI);自注意力机制;生成对抗网络(GAN);变分自编码器(VAE)

**DOI**:10.13568/j.cnki.651094.651316.2025.04.26.0001

**中图分类号**:R541.6;TP183 **文献标识码**:A **文章编号**:2096-7675(2026)01-0061-009

**引文格式**:张雅楠,张琳琳,郭渊博,毕雪华,赵楷. 一种用于心衰患者死亡率预测的数据多重插补方法[J]. 新疆大学学报(自然科学版中英文),2026,43(1):61-69.

**英文引文格式**:Zhang Yanan, Zhang Linlin, Guo Yuanbo, Bi Xuehua, Zhao Kai. A data multiple imputation method for mortality prediction in patients with heart failure[J]. Journal of Xinjiang University(Natural Science Edition in Chinese and English),2026,43(1):61-69.

# A Data Multiple Imputation Method for Mortality Prediction in Patients with Heart Failure

Zhang Yanan[1], Zhang Linlin[1,2], Guo Yuanbo[1], Bi Xuehua[3], Zhao Kai[4]

(1. School of Software, Xinjiang University, Urumqi Xinjiang 830091, China; 2. Center of Network and Information Technology, Xinjiang University, Urumqi Xinjiang 830046, China; 3. School of Medical Engineering and Technology, Xinjiang Medical University, Urumqi Xinjiang 830054, China; 4. School of Computer Science and Technology, Xinjiang University, Urumqi Xinjiang 830017, China)

**Abstract**: Aiming at the problem that the imperfect data collection mechanism of diagnosis and treatment leads to data loss and the poor quality of existing network feature extraction leads to the limited performance of clinical prediction models, a data multiple imputation method (Self-attention and Generative adversarial network based Mortality Prediction, SGMP) for mortality prediction of patients with heart failure is proposed. Firstly, the self-attention mechanism is used to dynamically fuse

* **收稿日期**:2025-04-26; **修回日期**:2025-11-14; **录用日期**:2025-11-16.

**基金项目**:国家自然科学基金“面向新疆棉花全产业链的数据要素权属确定关键技术研究”(62366052);新疆维吾尔自治区创新环境(人才、基地)建设专项(自然科学基金)联合基金“基于语义挖掘的大学生心理健康分析关键技术研究”(2022D01C427),“基于大模型微调和检索增强生成的高原病智能问答模型研究”(2024D01C126);四川省联合基金“高效区块链溯源与复杂查询技术研究”(25QYCX0103);省部共建中亚高发病成因与防治国家重点实验室开放课题“基于医学人工智能的心衰疾病进展模式的探究”(SKL-HIDCA-2022-23).

**作者简介**:张雅楠(2000—),女,硕士生,从事医学信息处理的研究,E-mail:107552305024@stu.xju.edu.cn.

† **通信作者**:赵楷(1976—),男,博士,副教授,主要从事海量数据分析处理及其应用、可信人工智能系统、软件和系统安全及生物信息处理的研究,E-mail:zhawkk@xju.edu.cn.

multiple sets of candidate estimates in the potential space of the variational autoencoder (VAE), and the adversarial training strategy of the generative adversarial network (GAN) is introduced to optimize the representation learning ability. Then, the candidate estimation results are effectively obtained according to the mask matrix, and the multiple interpolation of missing data is realized. Finally, the synthetic minority over-sampling technique (SMOTE) is used to enhance the data, and the multi-layer perceptron (MLP) is used to predict the mortality rate. Based on the diagnosis and treatment data of heart failure patients in a tertiary hospital in Xinjiang, the results show that: in the mortality prediction task, SGMP has significantly improved in multiple indicators compared with other models, and the area under the receiver operating characteristic curve reaches 0.902.

**Key words**: mortality prediction; multiple imputation (MI); self-attention mechanism; generative adversarial network (GAN); variational autoencoder (VAE)

## 0 引 言

心血管疾病(Cardiovascular Disease,CVD)的诊断已成为全球公共卫生领域的重大挑战. 世界卫生组织最新统计表明,心血管疾病每年导致的死亡人数达1 790万,占全球总死亡人数的32%,高于恶性肿瘤与呼吸系统疾病[1]. 心力衰竭(简称心衰)作为CVD的终末阶段,疾病防控形势尤为严峻. 据统计,全球每年有360万人被诊断为心衰,35%的患者在第一年内死亡,其余患者在5年内死亡[2]. 流行病学研究表明,过去15年,我国心衰患病率增长44%,新增确诊患者380万[3]. 心衰患者的死亡率因个体特征、合并症谱及医疗可及性等因素存在显著差异,范围介于5%~75%. 因此,如何基于心衰患者的特定情况构建精准预后模型,进而制定个体化干预方案、优化医疗资源配置,成为改善预后、降低患者医疗成本的关键.

作为临床决策支持系统的核心数据源,电子病历数据的完整性直接影响预后模型效能[4]. 然而,电子病历普遍存在数据缺失问题,主要归因于医疗设备故障、隐私保护策略下的信息屏蔽及未安排相关检查[5-6]. 数据缺失会引发特征空间分布偏移,导致预测模型准确率下降. 当前,数据插补技术主要分为单一插补(Single Imputation,SI)与多重插补(Multiple Imputation,MI). SI使用单组值(如均值插补)进行插补,虽计算高效,但生成数据分布与真实数据分布存在显著偏离[7]. 相比之下,MI通过模拟数据的后验分布,并融合从中采样的多组候选估计值生成最终插补结果,显著提升数据质量与预测可靠性[7].

近年来,深度学习在缺失数据插补中得到广泛应用,提高了下游任务的性能[8]. 基于自编码器(Autoencoder,AE)的方法,通过编码-解码架构学习数据潜在表征,在缺失值重建中展现出巨大潜力. Ma等[9]探索缺失值和非缺失值之间的非线性相关性,运用去噪自编码器补全数据. Kabir等[10]提出索引感知自编码器,增强缺失值的插补性能,提高疾病分类准确性. 作为AE的概率扩展模型,变分自编码器(Variational Autoencoder,VAE)通过重参数化技巧结合蒙特卡洛方法,利用多次采样数据支持多重插补. Chen等[11]通过正则化深度自编码器实现缺失值补全,并通过双分支VAE结合图推理,量化个性因果效应. 但VAE存在潜在变量过度正则化问题,表征学习能力受限,生成数据多样性不足. 生成对抗网络(Generative Adversarial Network,GAN)[12]通过生成器与判别器的对抗性优化框架,能够克服VAE在数据插补方面的不足. 生成器通过对抗训练生成逼近真实数据分布,判别器通过数据判别边界,提升数据的语义一致性. Zhao等[13]在GAN中嵌入自注意力机制,增强多变量时序数据特征相关性建模能力. 针对临床数据的高缺失率和不平衡分布问题,Dong等[14]提出对抗性插补网络(Generative Adversarial Imputation Nets,GAIN),在大型临床数据中进行插补,实现了更高的死亡率预测准确性.

尽管上述研究证实了深度学习方法的有效性,但在真实临床场景中仍面临挑战:多重插补融合机制粗粒度化导致数据填补偏差,VAE表征学习能力受限. 因此,本文提出一种用于心衰患者死亡率预测的数据多重插补方法SGMP,在潜在空间中引入自注意力机制,动态加权融合多组蒙特卡洛采样结果;同时结合GAN对抗训练思想,优化特征空间判别性,降低数据多重插补偏差,提升死亡率预测精度,为精准医疗决策提供高鲁棒性支持.

## 1 SGMP模型

在SGMP模型构建中,引入自注意力机制和GAN框架,显著提高数据信息捕捉与潜在表示学习能力. 以上改进可优化多重插补性能,提高数据质量,为死亡率预测提供更准确的输入,有效提高模型整体预测性能.

### 1.1　自注意力机制

自注意力机制凭借动态加权特性，有效融合全局信息，显著提升模型对特征复杂交互的建模能力．因此，本文引入自注意力机制，对多组插补候选估计值动态加权融合，实现融合机制的细粒度化，提高死亡率预测准确性．给定输入数据$X_{\text{input}} \in R^{n\times d}$，首先通过线性变换生成查询矩阵（$\boldsymbol{Q}$）、键矩阵（$\boldsymbol{K}$）和值矩阵（$\boldsymbol{V}$）．随后通过计算$\boldsymbol{Q}$与$\boldsymbol{K}$点积并归一化，得到注意力权重矩阵$\boldsymbol{A}$：

$$\boldsymbol{A}=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{T}}{\sqrt{d_k}}\right), \tag{1}$$

式中：$d_k$为键向量的维度．最终得到输出数据$X_{\text{output}}$：

$$X_{\text{output}}=\boldsymbol{A}\times\boldsymbol{V}. \tag{2}$$

### 1.2　生成对抗网络

GAN由生成器$G$和判别器$D$组成，通过对抗训练优化特征生成与判别能力，可生成高质量数据样本．因此，本文引入GAN有助于增强模型的潜在表示学习能力，提升数据质量和下游任务预测精度．$G$的目标是从先验分布中生成接近真实数据分布的样本，而$D$的目标是提高区分真实数据与生成数据的判别标准．因此，GAN目标函数为：

$$\min_{G}\max_{D} V(D,G)=E_{x\sim r(x)}\left[\log D(x)\right]+E_{a\sim f(a)}\left[\log(1-D(G(a)))\right], \tag{3}$$

式中：$x$为真实数据，采集于真实数据分布$r(\cdot)$；$a$为生成数据，采样于先验分布$f(\cdot)$；$E(\cdot)$用于计算期望．

### 1.3　SGMP模型结构

为训练高精度数据插补模型，本文首先在不含缺失值子集$X_{\text{obs}}$上对SGMP模型进行训练，进而将其应用于含缺失值子集$X_{\text{mis}}$以实现高质量数据插补．SGMP模型由三部分组成：潜在表示学习与多重融合、多重插补和死亡率预测，模型结构如图1所示．具体而言，在训练阶段，编码器在潜在空间中结合自注意力机制，对$X_{\text{obs}}$基于高斯分布采样的多个候选估计值进行融合，通过GAN框架不断优化VAE，以重建并生成最终插补结果．然后利用训练好的模型和记录缺失值位置的掩码矩阵对$X_{\text{mis}}$进行多重插补，并计算得到与$X_{\text{obs}}$拼接后的数据集$\tilde{X}$．最后将$\tilde{X}$输入到多层感知机（Multilayer Perceptron, MLP）中进行心衰患者死亡率预测．

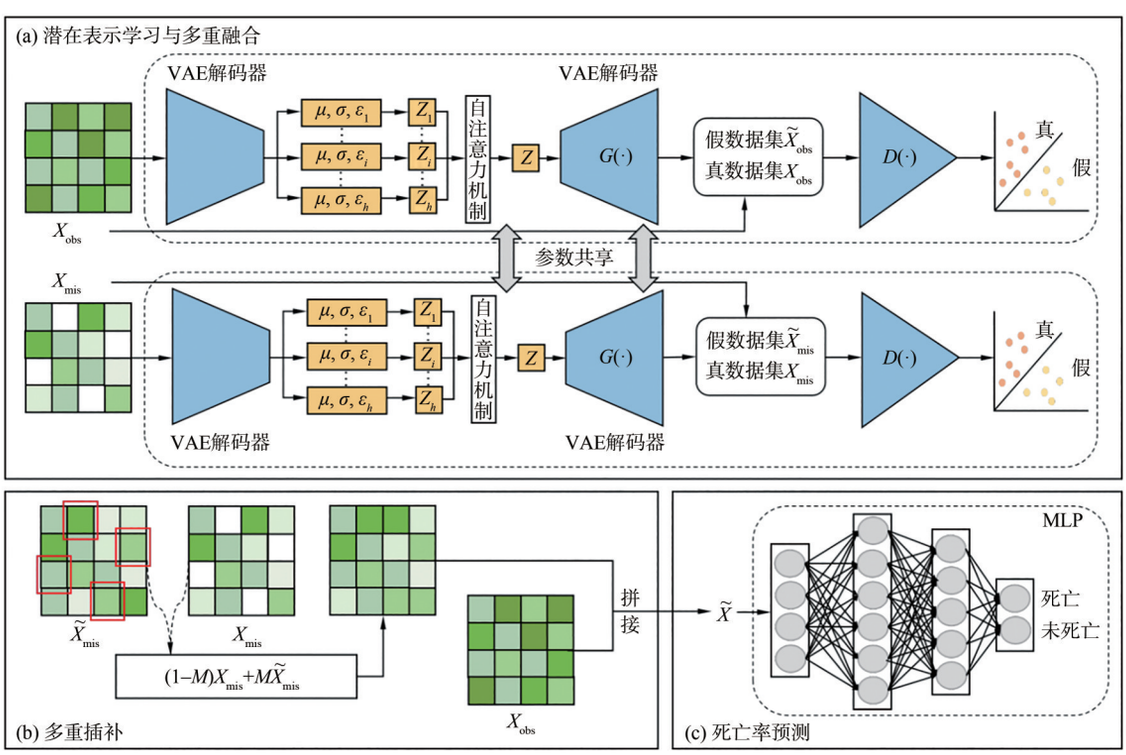


图 1　SGMP模型结构

**Figure 1　Structure of the SGMP model**

(a)潜在表示学习与多重融合. 给定$X_{\text{obs}}$,采用VAE的编码器网络$q_\phi(z|X_{\text{obs}})$,在潜在空间将其映射为高斯分布$N(\mu,\sigma^2)$. 为增加生成数据多样性并减少填补偏差,本文通过添加随机变量$\{\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_h\}$的重参数化法,基于高斯分布,采集$h$个候选估计值,记为$\{z_1,z_2,\cdots,z_h\}$,实现多重采样. 为实现多个候选估计值的融合机制细粒度化,进一步引入自注意力机制[15]. 输入候选估计值$z_i(i=1,2,\cdots,h)$,经过矩阵乘法生成对应的查询$\boldsymbol{Q}_i$、键$\boldsymbol{K}_i$和值$\boldsymbol{V}_i$向量:

$$\begin{cases}\boldsymbol{Q}_i=z_i\times\boldsymbol{W}^{Q},\\ \boldsymbol{K}_i=z_i\times\boldsymbol{W}^{K},\\ \boldsymbol{V}_i=z_i\times\boldsymbol{W}^{V},\end{cases}\tag{4}$$

式中:$\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$和$\boldsymbol{W}^{V}$为可学习参数矩阵. 通过式(5)实现多重融合,得到最终潜在变量$z$:

$$z=softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^{T}}{\sqrt{d_k}})\boldsymbol{V},\tag{5}$$

式中:$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$分别为查询、键和值矩阵. 解码器网络$p_\theta(X_{\text{obs}}|z)$将$z$映射到重建数据$\tilde{X}_{\text{obs}}$. VAE优化目标为证据下界:

$$L_{\text{VAE}}=E_{q_\phi(z|X_{\text{obs}})}[\log p_\theta(X_{\text{obs}}|z)]-D_{\text{KL}}(q_\phi(z|X_{\text{obs}})\|p_\theta(z)),\tag{6}$$

式中:第1项表示$z$到重构输入数据$X_{\text{obs}}$的损失,第2项表示$z$的近似后验$q_\phi(z|X_{\text{obs}})$和先验分布$p_\theta(z)$之间的KL散度. 为进一步增强潜在表示学习能力并提高重建数据质量,本文引入GAN框架,通过生成器$G$和判别器$D$的动态博弈机制实现协同优化[16]. 生成器采用VAE解码器,判别器则采用GAN的架构,优化公式为:

$$L_{\text{GAN}}=\min_G\max_D E_{X_{\text{obs}}\sim p(X_{\text{obs}})}\left[\log D(X_{\text{obs}})\right]+E_{z\sim p(z)}\left[\log(1-D(G(z)))\right].\tag{7}$$

(b)多重插补. 训练完成后,使用参数共享的模型对$X_{\text{mis}}$同样执行潜在表示学习与多重融合过程. 通过生成并融合多个候选估计值,得到优化后的候选估计结果$\tilde{X}_{\text{mis}}$,该结果用于实现多重插补. 此过程中,掩码矩阵$\boldsymbol{M}$用于标记缺失值位置,将$\tilde{X}_{\text{mis}}$数据集中对应位置的数值插补至$X_{\text{mis}}$中,从而完成多重插补. 最后再与$X_{\text{obs}}$拼接,表达式为:

$$\tilde{X}=(X_{\text{obs}},(1-\boldsymbol{M})X_{\text{mis}}+\boldsymbol{M}\tilde{X}_{\text{mis}}).\tag{8}$$

(c)死亡率预测. 本文采用MLP作为预测端,将死亡率预测任务建模为二分类任务(死亡/未死亡). 模型优化过程中,二元交叉熵函数公式为:

$$L_{\text{MP}}=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log(\hat{y}_i)+(1-y_i)\log(1-\hat{y}_i)\right],\tag{9}$$

式中:$y_i$和$\hat{y}_i$分别为真实标签和预测标签.

SGMP模型采用分阶段训练策略,依次优化多重插补损失$L_{\text{MI}}$与死亡率预测损失$L_{\text{MP}}$,以兼顾插补质量与分类性能. 由于多重插补关注对抗训练,权重分别为1.0和1.125. 待插补部分收敛后,固定模型参数,然后使用死亡率预测损失$L_{\text{MP}}$独立训练分类器. 该策略有效避免了多任务联合训练中的梯度冲突,确保数据插补与预测性能的协同提升. 总损失函数表达式为:

$$L=L_{\text{MI}}+L_{\text{MP}}=L_{\text{VAE}}+1.125L_{\text{GAN}}+L_{\text{MP}}.\tag{10}$$

## 2 实验与结果分析

### 2.1 数据集

本文真实数据来源于新疆某三甲医院心内科,收集2000年5月至2023年9月出院诊断第一诊断为心衰且年龄大于等于18岁的患者电子病历,涵盖住院期间人口统计学、病史记录、实验室检验指标及影像学检查结果特征. 为确保数据隐私安全,基于患者唯一标识符进行数据对齐并删除隐私信息. 数据清洗流程:将变量中Null值、零值和负值标记为缺失值,将连续特征异常值置为缺失,对离散型特征相似类别进行整合. 预处理后,共筛选出15 773例样本,构建真实心衰数据集EMR-HF,共105个临床特征,包括15个离散型特征、90个连续型特征. 该数据集中正样本(死亡)1 142例、负样本(存活)14 631例.

为验证模型泛化能力,基于MIMIC-III公共数据库[17],遵循上述筛选标准,通过25种心衰ICD-9诊断代码,

筛选前3诊断包含心衰的电子病历，经患者层级数据整合后获得5 011例样本．基于此构建的MIMIC-HF数据集共59个临床特征，包括2个离散型特征、57个连续型特征．该数据集中正样本（死亡）1 471例、负样本（存活）3 540例．

### 2.2　评价指标

评价指标用于定量衡量模型在心衰患者死亡率预测任务的性能表现．本文采用准确率$V_{ACC}$、精确率与召回率的调和平均数$F_1$、受试者工作特征曲线下面积$V_{AUC}$和精确率-召回率曲线下面积$V_{AUPR}$ 4个评价指标．

$V_{ACC}$用于衡量模型预测正确样本数量占总样本数量的比例，反映模型整体预测准确性，表达式为：

$$V_{ACC}=\frac{T_P+T_N}{T_P+T_N+F_P+F_N}, \tag{11}$$

式中：$T_P$、$T_N$、$F_P$和$F_N$分别为真正例、真负例、假正例和假负例．$F_1$为精确率与召回率的调和平均数，常在类别不平衡的情况下衡量模型的整体性能，表达式为：

$$F_1=2\times\frac{P\times R}{P+R}, \tag{12}$$

式中：$P$为精确率，即预测为正类别中实际为正类别的比例；$R$为召回率，即实际为正类别中被预测为正类别的比例．

$$P=\frac{T_P}{T_P+F_P}, \tag{13}$$

$$R=\frac{T_P}{T_P+F_N}. \tag{14}$$

$V_{AUC}$为受试者工作特征曲线下面积，该曲线根据模型正确预测为正的样本比例和模型错误预测为正的负样本比例绘制，表示模型区分正负类别的能力．$V_{AUPR}$为精确率-召回率曲线下面积，在类别不平衡数据集中，能更有效反映模型对正类别的预测性能．

### 2.3　实验结果

#### 2.3.1　候选估计值数量的敏感性分析

为评价候选估计值数量$h$对结果的影响，在真实数据集EMR-HF上进行敏感性分析．$h$取值分别为1、5、10、20、30、40和50，各评价指标随$h$变化的趋势如图2所示．当$h$为5时，模型的各项评价指标达到最佳性能，继续增加$h$时模型性能反而有所下降．因此，后续实验统一将$h$设定为5．

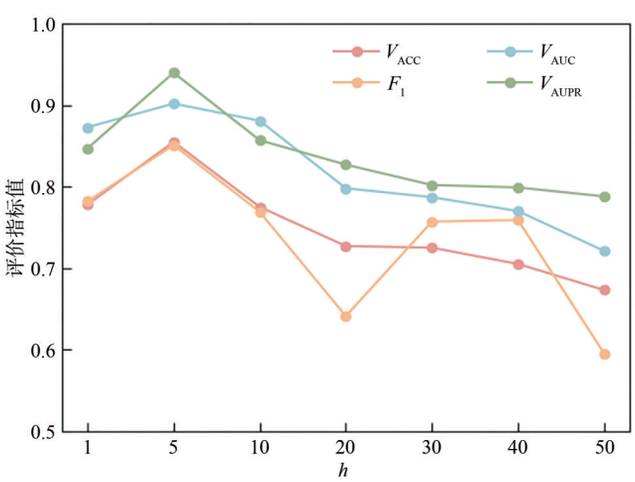


图 2　不同候选估计值数量下评价指标的变化

**Figure 2　Changes in evaluation metrics under different numbers of candidate estimates**

#### 2.3.2　死亡率预测结果分析

确定合适的分类方法对于死亡率预测具有重要意义．本文选取逻辑回归（Logistic Regression，LR）、K-近邻（K-Nearest Neighbor，KNN）、支持向量机（Support Vector Machine，SVM）和MLP 4种分类器在EMR-HF上

进行实验对比. 由于正负样本比例差异显著,采用五折交叉验证,其中:4份数据采用SMOTE[18]算法过采样,以平衡正负样本比例用于训练模型;1份数据用于验证模型. 最终取5次验证结果的平均值,综合评估各分类器性能. 分类器均使用过采后的数据进行训练与评估,在$V_{ACC}$、$F_1$、$V_{AUC}$和$V_{AUPR}$ 4个评价指标方面的表现如图3所示. MLP的$V_{ACC}$为0.783、$F_1$为0.779、$V_{AUPR}$为0.826,总体性能优于其他方法. 故将MLP作为心衰患者死亡率预测的分类器.

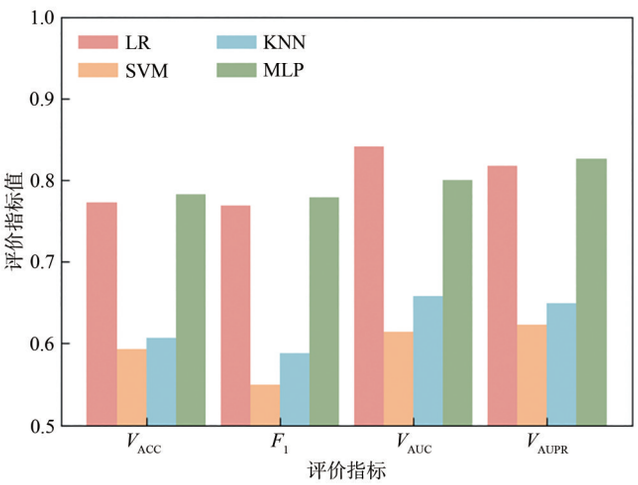


图 3 不同分类器下评价指标的对比

**Figure 3 Comparison of evaluation metrics across different classifiers**

为全面评估多重插补方法对预测任务的影响,本文在真实数据集EMR-HF和公共数据集MIMIC-HF上进行验证. 对比模型为其他6种插补方法,单值插补方法包括MissForest[19]、KNN[20]和SIVAE[21],多重插补方法包括MICE[22]、MIWAE[23]和PMIVAE[24]. 其中SIVAE是本文实现的基准模型,通过标准VAE的单次编码-解码流程实现插补,并移除SGMP中的自注意力融合模块与判别器. 为确保模型性能对比的公平性,所有方法均采用相同网络层数和神经元个数的MLP作为预测端,预测结果如表1所示.

表 1 在EMR-HF/MIMIC-HF上本文方法与其他方法预测结果

**Table 1 Prediction results of the proposed method and other methods on EMR-HF/MIMIC-HF**

| 数据集 | 方法 | $V_{ACC}$ | $F_1$ | $V_{AUC}$ | $V_{AUPR}$ |
|---|---|---|---|---|---|
| EMR-HF | MissForest[19] | 0.693±0.104 | 0.708±0.116 | 0.825±0.606 | 0.854±0.047 |
| | KNN[20] | 0.704±0.058 | 0.654±0.136 | 0.768±0.085 | 0.831±0.044 |
| | SIVAE[21] | 0.783±0.051 | 0.779±0.052 | 0.800±0.059 | 0.826±0.040 |
| | MICE[22] | 0.802±0.051 | 0.804±0.038 | 0.871±0.044 | 0.897±0.047 |
| | MIWAE[23] | 0.700±0.090 | 0.749±0.036 | 0.828±0.038 | 0.868±0.026 |
| | PMIVAE[24] | 0.498±0.030 | 0.198±0.256 | 0.556±0.011 | 0.606±0.013 |
| | SGMP | 0.855±0.105 | 0.851±0.102 | 0.902±0.081 | 0.940±0.050 |
| MIMIC-HF | MissForest[19] | 0.855±0.136 | 0.780±0.081 | 0.865±0.087 | 0.868±0.078 |
| | KNN[20] | 0.767±0.097 | 0.758±0.028 | 0.845±0.012 | 0.853±0.036 |
| | SIVAE[21] | 0.733±0.016 | 0.769±0.010 | 0.824±0.024 | 0.838±0.041 |
| | MICE[22] | 0.752±0.030 | 0.761±0.017 | 0.847±0.046 | 0.870±0.040 |
| | MIWAE[23] | 0.785±0.050 | 0.791±0.025 | 0.867±0.081 | 0.897±0.053 |
| | PMIVAE[24] | 0.522±0.027 | 0.394±0.037 | 0.544±0.006 | 0.625±0.002 |
| | SGMP | 0.857±0.024 | 0.856±0.020 | 0.914±0.013 | 0.916±0.014 |

在EMR-HF上,本文方法SGMP在$V_{ACC}$、$F_1$、$V_{AUC}$和$V_{AUPR}$指标上均优于其他方法,其中$V_{AUC}$和$V_{AUPR}$相较MICE分别提高了3.1%、4.3%,证明其可有效推断临床特征缺失值,并提高死亡率预测的准确性. 在MIMIC-

HF 上，SGMP 的 $V_{AUC}$ 和 $V_{AUPR}$ 分别达到 0.914 和 0.916，且 $V_{ACC}$、$F_1$ 和 $V_{AUC}$ 指标优于在 EMR-HF 上的表现，体现了模型的强泛化性．此外，PMIVAE 采用 VAE 框架结合蒙特卡洛方法实现部分多重插补，但在两个数据集上的效果均不如 SIVAE．而本文基于 VAE 改进的 SGMP 优于 PMIVAE 和 SIVAE，进一步验证了融合机制细粒度化与增强表征学习能力的有效性．

2.3.3　可解释性分析

全局解释通过计算特征的平均绝对 SHAP 值，量化各变量对预测结果的总体贡献度．如图 4 所示，特征重要性评分排名前 3 位的分别为 D-Dimer、NT-ProBNP 和 Crea，与文献[25-27]列明的核心预后生物标志物高度一致，验证了 SGMP 模型决策机制的临床合理性．而 Hct 与 PLT 的贡献度显著低于这些核心特征．

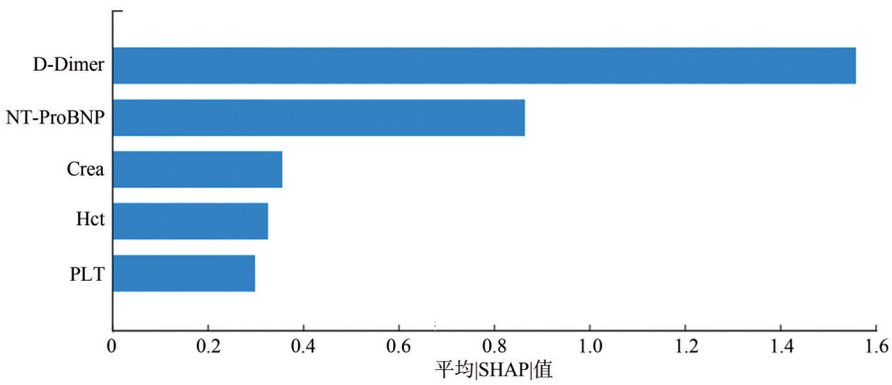


**图 4　SHAP 汇总图**

**Figure 4　SHAP summary plot**

局部解释基于 SHAP 分析量化，呈现个体样本中特征对预测的贡献方向与强度．如图 5 所示，蓝色特征代表降低患者发生死亡的概率，红色特征代表增加患者发生死亡的概率．D-Dimer 在真阴性和假阳性样本中 SHAP 值为负且绝对值较大，增加阳性预测风险；在假阴性和真阳性样本中 SHAP 值为正，降低阳性预测可能性．NT-ProBNP 在真阴性样本中数值高，支持正确预测；在真阳性样本中 SHAP 值高度正向，符合心衰诊断．Crea 在真阴性样本中 SHAP 值显著负向，增强阴性预测可能性；在真阳性样本中 SHAP 值正常，干扰阳性判定；在假阴性和假阳性样本中 SHAP 值中性，未提供有效信息，导致误判．

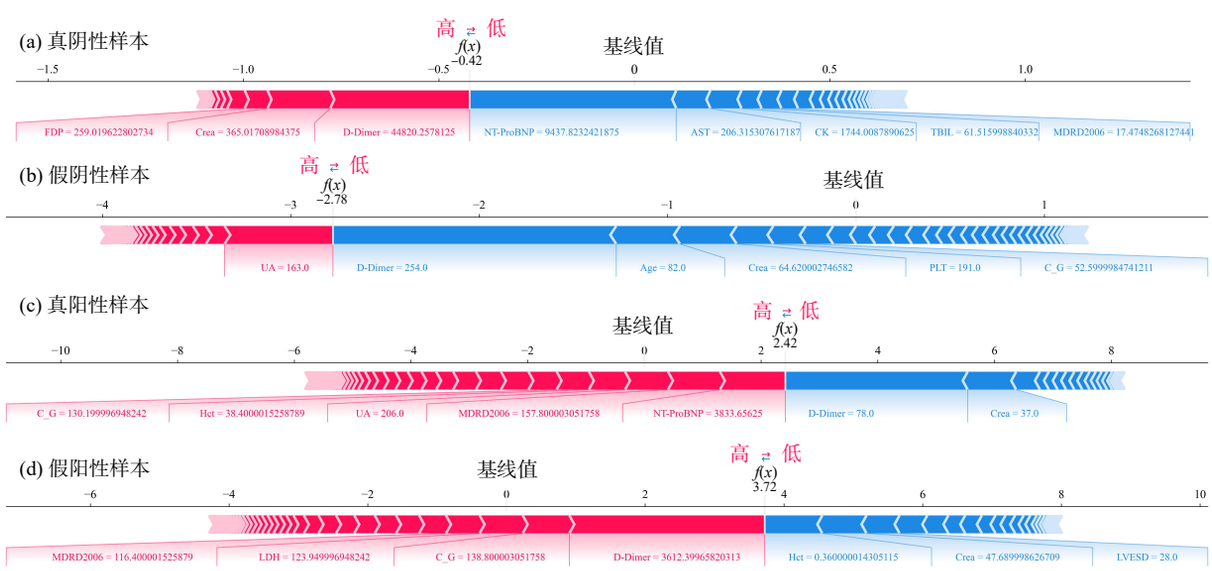


**图 5　四个代表性样本的 SHAP 力图**

**Figure 5　SHAP plots for four representative samples**

综上所述,进一步证明D-Dimer、NT-ProBNP和Crea是模型预测心衰患者死亡率的关键特征. 通过深入分析这些关键特征,临床医生可以更好理解模型的预测机制,从而在实际临床实践中更有效地应用模型,提高对心衰患者死亡风险的预测和管理能力.

## 3 结 论

本文提出一种用于心衰患者死亡率预测的数据多重插补方法SGMP,在VAE基础上引入自注意力机制,对多组候选估计值动态加权融合;结合GAN的对抗训练框架,提高生成表征学习质量,实现高质量数据多重插补,增强心衰患者死亡率预测的性能. 在真实数据集EMR-HF和公共数据集MIMIC-HF上的评估结果表明,本文方法相较于其他方法,具有良好的泛化性及预测性能,解决了电子病历数据缺失对临床预测模型的效能制约问题. 然而,目前的死亡率预测分类器仅采用MLP,未来将探索引入图神经网络等分类器,进一步提升死亡率预测的准确性和可靠性.

## 参考文献:


[1] Chen J T, Huang S, Zhang Y, et al. Congenital heart disease detection by pediatric electrocardiogram based deep learning integrated with human concepts[J]. Nature Communications, 2024, 15: 976.

[2] Chen J, Aronowitz P. Congestive heart failure[J]. Medical Clinics of North America, 2022, 106: 447-458.

[3] 付健. 基于MIMIC-III数据库的心衰患者死亡率预测模型研究[D]. 太原:太原理工大学,2021.
Fu J. Study of model for predicting mortality for heart failure patients based on MIMIC-III database[D]. Taiyuan: Taiyuan University of Technology, 2021. (in Chinese)

[4] Liu M X, Li S Q, Yuan H, et al. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques[J]. Artificial Intelligence in Medicine, 2023, 142: 102587.

[5] Bernardini M, Doinychko A, Romeo L, et al. A novel missing data imputation approach based on clinical conditional generative adversarial networks applied to EHR datasets[J]. Computers in Biology and Medicine, 2023, 163: 107188.

[6] Austin P C, White I R, Lee D S, et al. Missing data in clinical research: A tutorial on multiple imputation[J]. Canadian Journal of Cardiology, 2021, 37: 1322-1331.

[7] Zhao C, Su K J, Wu C, et al. Multi-scale variational autoencoder for imputation of missing values in untargeted metabolomics using whole-genome sequencing data[J]. Computers in Biology and Medicine, 2024, 179: 108813.

[8] Sun Y G, Li J, Xu Y F, et al. Deep learning versus conventional methods for missing data imputation: A review and comparative study[J]. Expert Systems with Applications, 2023, 227: 120201.

[9] Ma Q, Lee W C, Fu T Y, et al. MIDIA: Exploring denoising autoencoders for missing data imputation[J]. Data Mining and Knowledge Discovery, 2020, 34: 1859-1897.

[10] Kabir S, Farrokhvar L. Non-linear missing data imputation for healthcare data via index-aware autoencoders[J]. Health Care Management Science, 2022, 25: 484-497.

[11] Chen Z, Cao B, Edwards A, et al. A deep imputation and inference framework for estimating personalized and race-specific causal effects of genomic alterations on PSA[J]. Journal of Bioinformatics and Computational Biology, 2021, 19(4): 2150016.

[12] 何琨,陈振华,黄绍,等. 基于变分编码器与主成分分析的电磁对抗攻击[J]. 华中科技大学学报(自然科学版),2024,52(11):1-7.
He K, Chen Z H, Huang S, et al. Electromagnetic adversarial attack based on variational auto-encoder and principal component analysis[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2024, 52(11): 1-7. (in Chinese)

[13] Zhao J Q, Rong C T, Dang X, et al. QAR data imputation using generative adversarial network with self-attention mechanism[J]. Big Data Mining and Analytics, 2024, 7(1): 12-28.

[14] Dong W N, Fong D Y T, Yoon J S, et al. Generative adversarial networks for imputing missing data for big data clinical research[J]. BMC Medical Research Methodology, 2021, 21: 78.

[15] 施媛波. 变分自编码器和注意力机制的异常入侵检测方法[J]. 重庆邮电大学学报(自然科学版),2022,34(6):1071-1078.
Shi Y B. Anomaly intrusion detection method based on variational autoencoder and attention mechanism[J]. Journal of

Chongqing University of Posts and Telecommunications(Natural Science Edition),2022,34(6):1071-1078.(in Chinese)
[16] 粟佳,于洪. 基于条件生成对抗插补网络的双重判别器缺失值插补算法[J]. 计算机应用,2024,44(5):1423-1427.
Su J,Yu H. Missing value imputation algorithm using dual discriminator based on conditional generative adversarial imputation network[J]. Journal of Computer Applications,2024,44(5):1423-1427.(in Chinese)
[17] Johnson A E W,Pollard T J,Shen L,et al. Data descriptor:MIMIC-III,a freely accessible critical care database[J]. Scientific Data,2016,3:160035.
[18] Chawla N V,Bowyer K W,Hall L O,et al. SMOTE:Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research,2002,16:321-357.
[19] Stekhoven D J,Bühlmann P. MissForest:Non-parametric missing value imputation for mixed-type data[J]. Bioinformatics,2012,28(1):112-118.
[20] Beretta L,Santaniello A. Nearest neighbor imputation algorithms:A critical evaluation[J]. BMC Medical Informatics and Decision Making,2016,16(S3):74.
[21] Kingma D P,Welling M. Auto-encoding variational Bayes[PP/OL]. V11. arXiv(2022-12-10)[2025-01-15]. https://arxiv.org/pdf/1312.6114.
[22] van Buuren S,Groothuis-Oudshoorn K. MICE:Multivariate imputation by chained equations in R[J]. Journal of Statistical Software,2011,45(3):1-67.
[23] Mattei P A,Frellsen J. MIWAE:Deep generative modelling and imputation of incomplete data sets[PP/OL]. V2. arXiv(2019-02-04)[2025-01-15]. https://arxiv.org/pdf/1812.02633.
[24] Pereira R C,Abreu P H,Rodrigues P P. Partial multiple imputation with variational autoencoders:Tackling not at randomness in healthcare data[J]. IEEE Journal of Biomedical and Health Informatics,2022,26(8):4218-4227.
[25] Monzo L,Girerd N,Ferreira J P,et al. High risk of stroke in patients with worsening heart failure,reduced ejection fraction,coronary heart disease and sinus rhythm:Risk prediction score analysis from the COMMANDER-HF trial[J]. Journal of Cardiac Failure,2024,30:618-623.
[26] Fuery M A,Leifer E S,Samsky M D,et al. Prognostic impact of repeated NT-proBNP measurements in patients with heart failure with reduced ejection fraction[J]. JACC:Heart Failure,2024,12(3):479-487.
[27] Tolomeo P,Butt J H,Kondo T,et al. Independent prognostic importance of blood urea nitrogen to creatinine ratio in heart failure[J]. European Journal of Heart Failure,2024,26:245-256.

**责任编辑：** 张自强