DOI:10.16016/j.2097-0927.202107161

基于 XGBoost 的 COVID-19 患者重症风险早期预测模型的建立与评价

王 铭[1,3*],程振豪[1,3*],胡 苗[2*],唐铭成[1,3],徐福民[1],王 莉[2],粘永健[2],刘凯军[1,4] 400042 重庆,陆军特色医学中心消化内科[1];400038 重庆,陆军军医大学(第三军医大学):生物医学工程与影像医学系[2],基础医学院学员五大队[3];430010 武汉,武汉市火神山医院感染一科[4]

[摘要] **目的** 利用新型冠状病毒病(corona virus disease 2019,COVID-19)患者的临床特征数据构建 XGBoost 预测模型,并评价预测模型对 COVID-19 患者重症进展风险早期预测的效能。**方法** 对 2020 年 2 月 10 日至 4 月 5 日火神山医院病案系统内经实验室确诊的 COVID-19 患者进行筛选,共收集 347 例有完整医疗信息和实验室检查结果的患者数据。首先筛选出 21 个具有显著性差异的指标作为训练模型的输入特征;对构建的 XGBoost 模型进行贝叶斯优化以调整参数,并基于特征重要性筛选出最优特征组合;进一步分析各特征数值大小对预测结果的正负影响,利用 SHAP(SHapley Additive exPlanation)对各特征重要性进行量化和归因;对 XGBoost 预测模型进行性能评价,并将其与其他机器学习方法进行对比,讨论其优势所在。**结果** 本研究选取 21 个重症组与非重症组差异显著的特征进行训练和验证。在 K 最邻近(k-nearest neighbor,KNN)模型中具有 10 个特征的最优子集获得了验证集中 4 个模型中曲线下面积(area under curve,AUC)值的最高值。年龄、脉率、白细胞计数、中性粒细胞计数、C-反应蛋白、总胆红素、肌酐、D-二聚体(D-Dimer)越高,疾病重症风险越高;淋巴细胞计数、白蛋白水平越低,疾病重症风险越高。XGBoost 与支持向量机的预测性能优于其他机器学习方法(在测试集上 AUC 值分别为 0.942 0、0.959 4),其中 XGBoost 训练速度明显更优。**结论** 基于 XGBoost 成功建立了预测模型,以最优特征子集实现了对 COVID-19 患者重症进展风险的早期预测。

[关键词] COVID-19;重症风险;预测模型;XGBoost;SHAP

[中图法分类号] R319; R512.99 [文献标志码] A

Early prediction model for disease progression of COVID-19 patients based on XGBoost: establishment and evaluation

WANG Ming[1,3], CHENG Zhenhao[1,3], HU Miao[2], TANG Mingcheng[1,3], XU Fumin[1], WANG Li[2], NIAN Yongjian[2], LIU Kaijun[1,4] [1]Department of Gastroenterology, Army Medical Center of PLA, Chongqing, 400042; [2]Faculty of Biomedical Engineering and Imaging Medicine, [3]Students' Team 5, College of Basic Medical Sciences, Army Medical University (Third Military Medical University), Chongqing, 400038; [4]First Department of Infectious Diseases, Wuhan Huoshenshan Hospital, Wuhan, Hubei Province, 430010, China

[**Abstract**] **Objective** To construct an XGBoost prediction model to predict disease severity of COVID-19 based on clinical characteristics dataset of COVID-19 patients. **Methods** A total of 347 laboratory-confirmed COVID-19 patients with complete medical information admitted from Feb 10 to April 5, 2020 were screened from the medical record system of Huoshenshan Hospital. Firstly, 21 features with significant differences were screened out as input features for the training model. Bayesian optimization was performed on the constructed XGBoost model to adjust the parameters, and the optimal combination of features was filtered based on feature importance. To further analyze the positive and negative effects of the numerical

* 同等贡献者

[通信作者] 刘凯军,E-mail:kliu_tmmu@126.com

2097-0927/Copyright © 2022, 陆军军医大学(第三军医大学)。这篇开放获取文章遵循 CC BY 许可协议(https://creativecommons.org/licenses/by/4.0/)

size of each feature on the prediction results, each feature importance was quantified and attributed by using SHapley Additive exPlanations (SHAP). Finally, the performance of the XGBoost prediction model was evaluated, and the model was compared and discussed with other machine learning methods, including support vector machine (SVM), naïve Bayes (NB), logical regression (LR), and k-nearest neighbors (KNN). **Results** In this study, 21 features with significant differences between the severe and non-severe groups were selected for training and validation. The optimal subset with 10 features in the k-nearest neighbor model obtained the highest value of area under curve (AUC) among the 4 models in the validation set. XGBoost and support vector machine were better than other machine learning methods in terms of prediction performance (AUC: 0.942 0, and 0.959 4 on the test set, respectively), and the training speed of XGBoost was significantly faster. **Conclusion** A prediction model based on XGBoost is successfully built to achieve early prediction of disease severity of COVID-19 patients.

[**Key words**] COVID-19; risk for severe illners; prediction model; XGBoost; SHAP

WANG Ming, CHENG Zhenhao and HU Miao contributed equally to the article. Corresponding author: LIU Kaijun, E-mail: kliu_tmmu@126.com

2019 年,新型冠状病毒病(corona virus disease 2019,COVID-19)肆虐全球。截至 2021 年 7 月 23 日,全球感染者约有 1.92 亿人,死亡病例达到 413.65 万人[1]。目前,新型冠状病毒仍在全球多地散发和流行,尤其是 Delta 变异株的出现增加了 COVID-19 再次大流行的风险[2]。新冠肺炎症状多样,大部分患者症状较轻,以发热、干咳、乏力等为主要表现,约 15.7%~32.0%患者会发展为重症[3]。重症病例多在 1 周后出现呼吸困难,严重者快速进展为急性呼吸窘迫综合征、脓毒症休克、难以纠正的代谢性酸中毒和凝血功能障碍及多器官功能衰竭等[4-5]。研究表明,重症患者预后极差且病死率显著升高[4,6-7]。因此,早期发现疾病进展风险高的患者,并提供重症监护和适当的干预,如优化危重症新冠肺炎患者呼吸支持和管理,将有助于提高重症患者的救治成功率且能合理分配医疗资源。新冠肺炎患者重症化的危险因素包括高龄、合并其他疾病以及 D-二聚体(D-Dimer)的水平升高等[8]。目前,仍然没有治疗新冠肺炎的有效药物,早期识别有重症化风险的患者对于疫情控制,疾病救治有着重要的临床意义。

机器学习(machine learning,ML)是人工智能的核心,其数据处理、归纳、综合能力远优于其他的统计学方法。机器学习方法在临床结局预测和危险因素评估等方面具有明显优势。近年来,一种基于回归树的提升算法——极限梯度提升(extreme gradient boosting,XGBoost)凭借训练时间短、精度高的优势被各界广泛应用。本研究在前期采用机器学习预测新冠肺炎重症化趋势的基础上[9],结合 COVID-19 患者的大量临床特征,构建 XGBoost 预测模型,从而对患者重症风险做出有效预测,并评价模型的准确性及临床价值,为临床决策提供重要的技术支持,为新冠肺炎防治提供可靠的预测模型。

1 材料与方法

1.1 数据预处理

1.1.1 训练集和测试集 本研究对 2020 年 2 月 10 日至 4 月 5 日火神山医院实验室确诊的 COVID-19 患者进行筛查并收集数据。在排除无完整病历信息的患者后,共收集 347 例有完整医疗信息和实验室检查结果患者的临床数据。将其中 269 例患者的数据作为训练集和验证集(重症 101 例,非重症 168 例),剩余 78 例患者的数据作为测试集(重症 30 例,非重症 48 例)。本研究所采集的临床数据均为患者入院后 6 小时内采集的数据。COVID-19 的诊断标准遵循 2020 年 WHO 指南[10]。重症 COVID-19 病例被定义为发热加上呼吸频率>30/min、严重呼吸窘迫或血氧分压<93%。所有重症病例均由非重症病例发展而来,临床数据采集时未发展为重症病例。患者无重症趋势、无需吸氧则被定义为非重症病例。所有非重症病例均稳定出院。鼻拭子和咽拭子标本被用于 SARS-CoV2 病毒 RT-PCR 检测。本研究被武汉火神山医院伦理委员会批准同意(HSSLL024),并遵循赫尔辛基宣言执行。

首先将训练集与验证集的所有样本数据分成重症组和非重症组,其中由轻症发展为重症标注为 1(阳性),而未发展为重症标注为 0(阴性)。利用 SPSS 软件 t 检验、非参数检验或 χ^2 检验筛选出两组间有显著性差异的指标进行后续的机器学习,在保证预测性能的前提下,显著降低后续预测过程的复杂度。从 29 个临床指标中筛选出 21 项两组间有显著差异指标,如表 1 所示。

表 1　COVID-19 患者重症组与非重症组有显著差异的 21 项指标中英文名及其缩写

序号	中文名	英文名	缩写
1	年龄	age	Age
2	基础疾病	comorbidity	Comorbidity *
3	呼吸频率	respiratory rate	RR
4	脉率	pulse rate	PR
5	白细胞计数	white cell count	WBC
6	中性粒细胞计数	neutrophil count	NEC
7	淋巴细胞计数	lymphocyte count	LYMPH
8	血红蛋白	hemoglobin	HGB
9	红细胞压积	hematokrit	HCT
10	血小板计数	platelets	PLT
11	C-反应蛋白	c-reactive protein	CRP
12	白蛋白	albumin	ALB
13	碱性磷酸酶	alkaline phosphatase	ALP
14	总胆红素	total bilirubin	TBIL
15	血尿素氮	blood urea nitrogen	BUN
16	肌酐	creatinine	Creatinine
17	血糖	glucose	Glucose
18	血清钙	calcium	Calcium
19	乳酸脱氢酶	lactic dehydrogenase	LDH
20	α-羟丁酸脱氢酶	alpha-hydroxybutyric dehydrogenase	α-HBDH
21	D-二聚体	D-Dimer	D-D

*包括合并糖尿病、高血压、冠心病、慢性阻塞性肺疾病、肿瘤、慢性肾病等 1 种或多种

1.1.2　数据归一化　　由于本研究采集的数据来自不同的医学仪器,存在量纲不一致,有的数据之间的差别可能很大,为了消除指标间的量纲和取值范围差异的影响,减少数据的冗余性和不确定性,需要对各指标数据进行归一化处理,以保证计算结果的科学、准确性。本文采用 min-max 标准化方法将数值映射到[-1,1]区间,计算方法如下:

$$X=2\left(\frac{X-X_{min}}{X_{max}-X_{min}}\right)-1$$

其中 X_{max} 为样本数据的最大值;X_{min} 为样本数据的最小值。

1.2　COVID-19 患者疾病重症化风险的早期预测模型构建

1.2.1　总体研究方案　　总体研究方案如图 1 所示。首先整理火神山医院 COVID-19 患者的样本数据,总的样本数据集分为训练集、验证集以及测试集。利用训练集和验证集上的数据构建和训练 XGBoost 预测模型,通过贝叶斯优化调整参数,并基于特征重要性排序进行指标集的优选,利用 SHAP(SHapley Additive exPlanation),一种附加的特征归因方法解释模型。为了评估 XGBoost 模型的预测性能,我们在测试集上利用受试者工作曲线(receiver operating characteristic curve,ROC)、曲线下面积(area under curve,AUC)值、敏感性和特异性等指标来评估所建立模型的预测性能,并与支持向量机(support vector machine,SVM)、朴素贝叶斯(naïve Bayes,NB)、逻辑回归(Logical regression,LR)、K 最邻近(k-nearest neighbor,KNN)等 4 种机器学习方法进行比较。

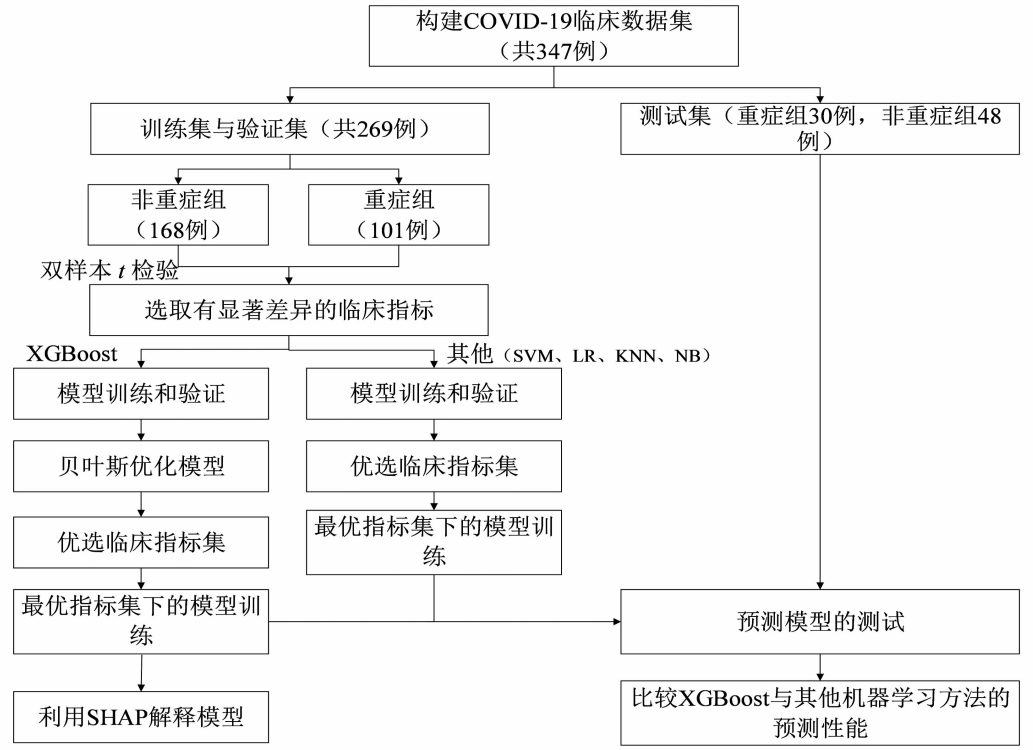


图 1　COVID-19 患者重症化预测模型构建研究方案流程图

1.2.2 模型选择与优化 本研究希望通过有限的训练样本去拟合一个模型来模拟理想模型。为降低模型的预测误差,大量参数被加入模型,使模型变得更加复杂。因此模型在训练样本中的偏差减小,但方差变大,导致模型的泛化能力较差;使新的预测样本预测效能下降,即模型出现了过拟合。为了避免过拟合,本研究减少了模型复杂度,并增加一些正则项来限制模型在训练过程中对训练数据的过度依赖,提高模型的稳定性,降低方差。同时,模型偏差也会增大,若模型拟合不充分,会导致模型预测能力下降,出现欠拟合。因此,我们需要在二者之间找到一个平衡,即偏差-方差权衡(bias-variance tradeoff)。在模型最佳的平衡位置,随着模型复杂度的变化,偏差的增加和方差的减少相等。如果超过了该最佳平衡点,模型会出现过拟合;若未达到该平衡点,说明模型欠拟合。

1.2.3 基于特征重要性的指标集优选 机器学习中特征越多,给出的信息就越多,识别准确性会得到提升。但特征多,计算复杂度增加,探索的空间就大,训练数据在全体特征向量中就会显得稀疏,影响相似性判断。更重要的是,如果有对分类无益的特征,反而可能干扰学习效果。因此,特征并非越多越好,获得有效的特征是研究的关键。

本研究在最优模型的基础之上,利用 feature importances 获取模型的特征重要性评分,并将其作为阈值进行特征选择,相同评分无须重复进行,因此要对特征重要性进行去重、排序的处理。

2 结果

2.1 XGBoost 模型预测结果

XGBoost 是一个优化的分布式梯度增强库,旨在实现高效、灵活和便捷。它在 Gradient Boosting 框架下实现机器学习算法。对于得到的 21 个指标,如表 2 所示,当特征数量为 10 个时,AUC 最高为 0.961,通过图 2 的 ROC 曲线也可以得到相同结果。基于特征重要性进行特征选择后得出 XGBoost 模型下的最优特征组合为 Age、PR、WBC、NEC、LYMPH、CRP、ALB、TBIL、Creatinine、D-Dimer,相应的 AUC 值为 0.961 0,灵敏度为 0.857 1,特异性为 0.939 4,F1 值为 0.878 0。

2.2 XGBoost 模型评价

将选出的 10 个最优特征集合在训练集上进行模型训练,并在测试集上进行预测,预测结果较好。AUC 值为 0.942 0,灵敏度为 0.900 0,特异性为 0.895 8,F1 值为 0.871 0。ROC 曲线如图 3 所示。

表 2 COVID-19 患者选择不同临床指标数量对应 AUC 值

特征数量	Age	Comorbidity	RR	PR	WBC	NEC	LYMPH	HGB	HCT	PLT	CRP	ALB	ALP	TBIL	BUN	Creatinine	Glucose	Calcium	LDH	α-HBDH	D-Dimer	AUC
1											√											0.818
2											√										√	0.812
3						√					√			√								0.818
4						√	√				√										√	0.812
5	√					√	√				√										√	0.812
6	√				√	√	√				√										√	0.836
7	√				√	√	√				√			√							√	0.853
8	√				√	√	√				√	√		√							√	0.892
9	√			√	√	√	√				√	√		√							√	0.946
10	√			√	√	√	√				√	√		√					√		√	0.961
11	√			√	√	√	√				√	√		√		√			√		√	0.946
12	√			√	√	√	√				√	√		√		√		√	√		√	0.907
13	√	√		√	√	√	√				√	√		√		√		√	√		√	0.892
14	√	√	√	√	√	√	√				√	√		√		√		√	√		√	0.892
15	√	√	√	√	√	√	√	√			√	√		√		√		√	√		√	0.907
16	√	√	√	√	√	√	√	√	√		√	√		√		√		√	√		√	0.877
17	√	√	√	√	√	√	√	√	√	√	√	√		√		√		√	√		√	0.907
18	√	√	√	√	√	√	√	√	√	√	√	√		√	√	√		√	√		√	0.922
19	√	√	√	√	√	√	√	√	√	√	√	√		√	√	√		√	√	√	√	0.907
20	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√		√	√	√	√	0.898
21	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	0.907

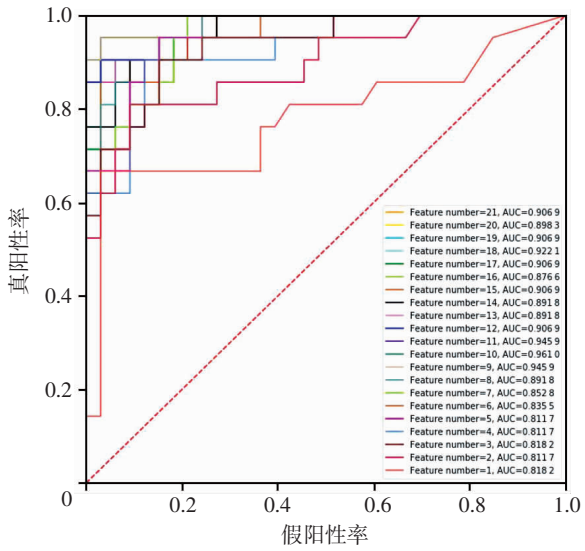


图 2 COVID-19 患者不同临床特征数量下的 ROC 曲线

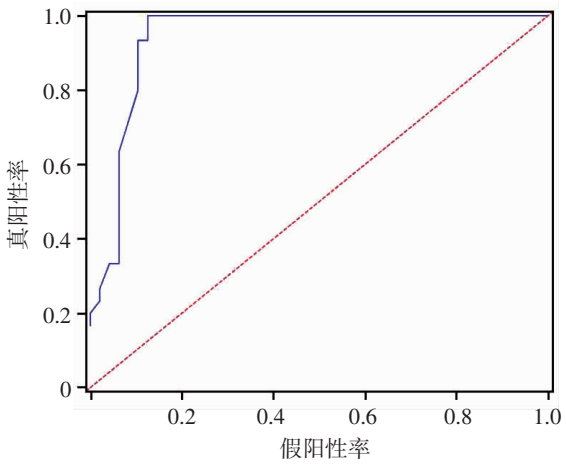


图 3 10 个最优特征在测试集上训练的 ROC 曲线

2.3 利用 SHAP 解释 XGBoost 模型

SHAP 是由 Shapley value 启发的可加性解释模型。对于每个预测样本,模型都产生一个预测值,SHAP value 就是该样本中每个特征所分配到的数值。已经得到最优预测模型及其所需最优指标集,来探究每个特征对结果的影响,以及特征值大小与重症风险之间的关系。如图 4 所示:对于该阴性样本,CRP、TBIL、Age 等数值较小且为蓝色,象征着这些特征将该样本的 SHAP 值拉低,对结果起负向作用。此外,D-Dimer 的蓝条最宽,象征着其对结果的影响最大。

如图 5 所示,对于该阳性样本,CRP、NEC、TBIL 等数值较大且为红色,象征着这些特征对结果起正向作用。此外,CRP 的红条最宽,象征着其对结果的影响最大。如图 6 所示,接着对各特征进行总体分析,按各特征的重要程度进行排序。

通过 shap. summary_plot 对特征进行整体的可视化,还可以分析各特征对预测结果的正负影响。如图 7 所示每一行代表一个特征,横坐标为 SHAP 值。一个点代表一个样本,颜色越红说明特征本身数值越大,颜色越蓝说明特征本身数值越小。可以直观地看出 CRP 是一个很重要的特征,基本上是与疾病进展风险成正相关的。D-Dimer、TBIL 等也会明显影响疾病进展,蓝色点主要集中在 SHAP<0 的区域,可见 D-Dimer、TBIL 小会降低疾病进展风险。

SHAP 也提供了部分依赖图的功能,与其他的部分依赖图不同的是,这里纵坐标不是目标变量 y 的数值而是 SHAP 值。通过 shap. dependence_plot 绘制(图 8),显示 D-Dimer、TBIL、CRP、Age 等与 SHAP 值呈正相关,即数值越大,疾病进展风险越高;LYMPH、ALB 与 SHAP 值呈负相关,即数值越小,疾病进展风险越高。

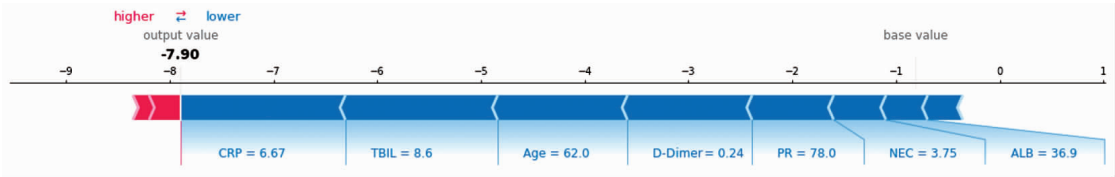


图 4 COVID-19 非重症组患者不同临床特征的 SHAP 值

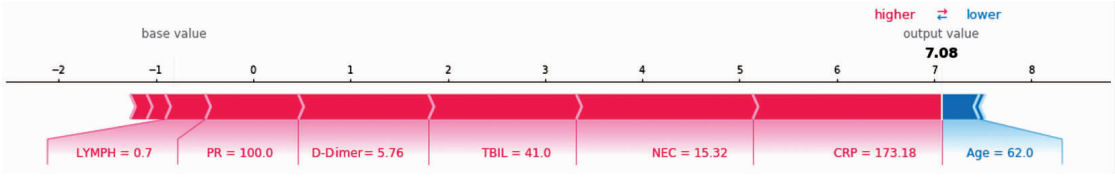


图 5 COVID-19 重症组患者不同临床特征的 SHAP 值

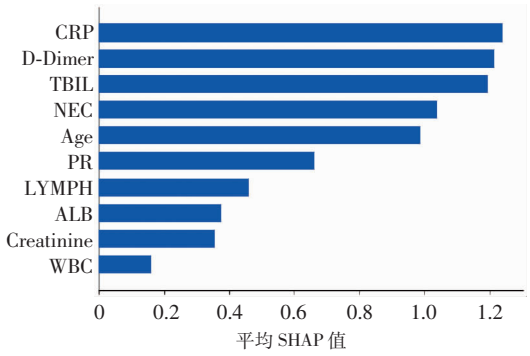


图 6 COVID-19 患者临床指标特征对重症化重要性排序

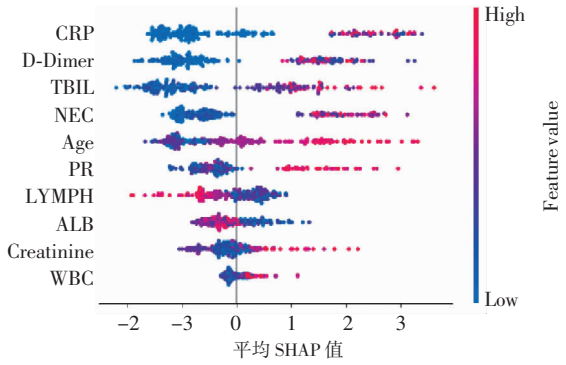


图 7 COVID-19 患者临床指标 SHAP 值可视化图

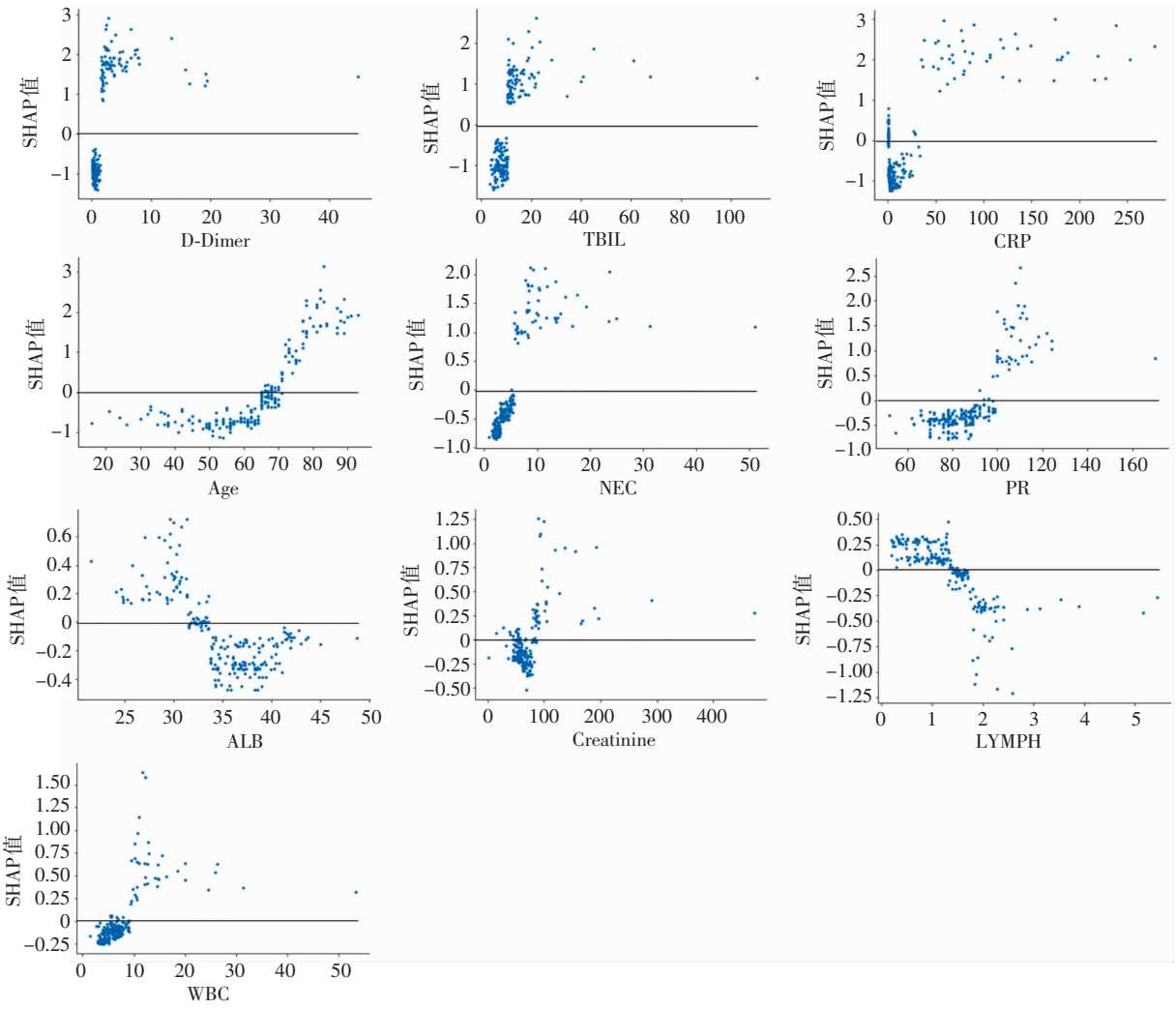


图 8 各指标 SHAP 值的分布图

2.4 XGBoost 与各机器学习方法预测性能比较

将各预测模型在测试集上进行预测,比较基于最优特征子集的 XGBoost、SVM、LR、KNN 和 NB 的预测性能。上述五种模型的敏感度、特异性、F1 值如表 3 所示。由结果可以发现,预测性能最好的是 SVM 预测模型,AUC 为 0.959 4,灵敏度为 0.900 0,特异性为 0.916 7,F1 值为 0.885 2。XGBoost 预测性能较好,AUC 为 0.942 0,灵敏度为 0.900 0,特异性为 0.895 8,F1 值为 0.871 0。预测性能最差的是逻辑回归预测模型,AUC 为 0.905 9,灵敏度为 0.900 0,特异性为 0.812 5,F1 值为 0.818 2。

通过观察最佳组合可以发现,PR、WBC、CRP、D-Dimer 在所有的最佳特征组合中都有出现,符合临床研究结果。此外,Age、RR、ALB 也是出现频率较高

的特征指标,同样符合临床研究结果。通过 SHAP 对 XGBoost 模型的解释,得出结论:D-Dimer、TBIL、CRP、Age、PR、Creatinine、WBC 数值越大,疾病进展风险越高;LYMPH、ALB 值越大,疾病进展风险越小,符合临床研究结果。这可以说明本研究的结果具有临床意义,可以作为临床医生诊断鉴定的一个辅助。

通过比较基于最优特征子集的 XGBoost 与其他 4 种机器学习方法(SVM、LR、KNN、NB)的预测性能(图 9),根据五种机器学习方法计算的结果可知,XGBoost 和 SVM 算法在 COVID-19 患者重症风险的预测上都具有较高的性能,在训练速度方面,XGBoost 明显优于 SVM,综合考虑模型各方面因素,我们推荐使用 XGBoost 算法作为预测模型的算法,Age、PR、WBC、NEC、LYMPH、CRP、ALB、TBIL、Creatinine、D-Dimer 作为预测所需指标,对早期 COVID-19 患者进展为重症的风险进行预测。

表 3 最优特征子集下不同算法测试集预测结果

算法	AUC	Sn	Sp	F1
XGBoost	0.942 0	0.900 0	0.895 8	0.871 0
SVM	0.959 4	0.900 0	0.916 7	0.885 2
NB	0.917 7	0.866 7	0.916 7	0.866 7
LR	0.905 9	0.900 0	0.812 5	0.818 2
KNN	0.913 9	0.800 0	0.916 7	0.827 6

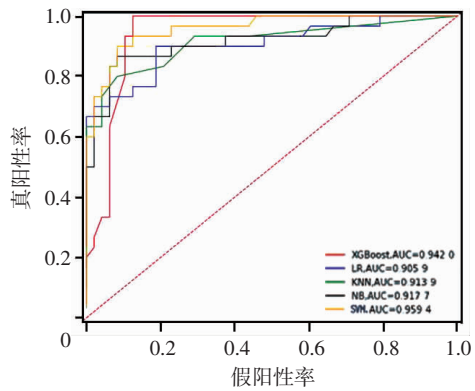


图 9 基于最优特征子集各模型测试集 ROC 曲线

3 讨论

虽然新冠肺炎在我国得到了有效的控制,但在全球范围内仍然在散发和流行。对于医疗水平薄弱的国家和地区,新冠疫情仍然会带来沉重的医疗和经济负担。因此,建立一种更简单、更准确、更具特异性的早期预测疾病进展风险模型,将会为临床决策的制定提供更客观准确的依据,有助于提高重症患者的救治率、降低重症患者病死率。

3.1 预测模型的临床特征及其意义

本研究纳入了 347 例 COVID-19 患者的临床特征包括血常规、生化、炎症标志物、凝血功能等,进行 t 检验或非参数检验以及数据的归一化处理,找到有意义的临床特征。运用 XGBoost 建立相应的预测模型,利用贝叶斯优化调整参数,得到最优模型。基于特征重要性筛选得到最优特征集合并对预测模型进行训练验证,得到相应的预测模型;将模型应用于测试集,对预测模型进行评价,获得最终预测后的 ROC 曲线,并计算 AUC、特异性、灵敏度等参数。利用 SHAP 对 XGBoost 模型进行解释,分析各特征对预测结果的贡献以及正负向影响。同时,利用其他 4 种机器学习方法(SVM、NB、LR、KNN)构建预测模型,在各模型下利用 k 折交叉验证筛选出最优指标集合,并将预测模型应用于测试集,比较 XGBoost 与其他机器学习方法的预测性能。通过观察最佳组合可以发现,PR、WBC、CRP、D-Dimer 是最常用的特征指标,与临床研究结果相符合。本研究最终得出的最优特征组合具有临床意义,对于早期 COVID-19 患者疾病重症风险的预测具有良好的表现。

本研究筛选出预测重症化的临床特征与既往临床回顾性研究结果相似,脉率增快、白细胞、C-反应蛋白、D-Dimer 升高与患者重症化趋势呈正相关。JI 等[11]开发了一种基于这些特征的风险因素评分系统(CALL)用来预测疾病进展。然而,纳入的病例很少,而且模型的可靠性有待进一步确认。ZHOU 等[12]发现病情加重组和非加重组之间白细胞计数、CRP、白蛋白、LDH 及 D-Dimer 水平没有明显差异。他们使用二元逻辑回归模型发现,COVID-19 患者的总淋巴细胞计数是与疾病进展相关的危险因素。但是,该研究仅纳入 17 名患者,淋巴细胞计数并不能反映疾病的进展。ZHOU 等[8]发现,高龄和 D-Dimer 水平升高有助于临床医生在早期发现预后不良的患者。与本研究一致,D-Dimer 水平是最佳特征子集中的重要特征。D-Dimer 水平升高与疾病活动和炎症有关,主要包括静脉血栓栓塞、败血症或癌症。对死亡患者的回顾性研究还表明,死亡患者的 D-Dimer 明显高于康复患者[13]。因此,监测 D-Dimer 水平可以帮助临床医生找出疾病恶化的高风险患者。D-Dimer 水平高的患者可给予抗凝治疗,以防止疾病进展。大多数严重 COVID-19 患者的白蛋白水平显著降低,并在疾病进展过程中持续降低。低白蛋白血症与住院患者的不良临床结局相关。重症患者的低白蛋白血症主要是由于营养摄入不足和过度消耗所致。本研究中胆红素升高对于新冠肺炎重症化风险有预测作用。有研究表明,新冠肺炎肝损伤率大概在 14.8%~53%,重症组肝损伤的比例显著高于轻症组,同时肝功能损伤是 COVID-19 患者临床结局较差的独立危险因素[14-16]。

3.2 模型的临床应用价值

本研究基于XGBoost构建出COVID-19患者疾病重症风险的早期预测模型,能够为患者提供个体化的预测概率。经过内部验证,模型的AUC值高达0.942 0。该模型应用简便,只需输入所需计算的患者10个变量,便可快速得到其重症化的概率。如果患者进展为重症疾病的概率较低,医生可以选择对其进行病情监测,若概率较高,则应尽快进行积极治疗或转入ICU病房。

本研究通过SHAP分析了各个指标数值大小对预测结果的影响,可与临床医生的诊断鉴定相互验证和促进;此外,将XGBoost和其他的4种机器学习方法(SVM、NB、LR、KNN)进行比较,发现不同分类模型作用于同一疾病时,需要的特征指标也不一样。XGBoost和SVM算法在COVID-19患者重症风险的预测上都具有较高的性能,SVM算法AUC较高,在训练速度方面,XGBoost优于SVM。XGBoost是boosting算法的其中一种。Boosting算法的思想是将许多弱分类器集成在一起形成一个强分类器。因为XGBoost是一种提升树模型,所以它是将许多树模型集成在一起,形成一个很强的分类器。而所用到的树模型则是CART回归树模型。它具有速度快、效果好、能处理大规模数据、支持多种语言、支持自定义损失函数的优点。从统计学和医学角度看,还有一些方面需要继续研究:患者的样本量比较小,阴性样本和阳性样本数量不平衡;采集的患者数据均来自火神山医院,预测模型只进行了内部验证,没有在其他医院进行多中心验证,无法评判其泛化能力;本研究是回顾性研究,在后续的COVID-19重症患者临床资料数据库的建立上,可为前瞻性研究提供相应支持。在后续研究中可开发一应用于临床的预测软件,为COVID-19重症化早期预测提供有效工具。

参考文献:

[1] WHO. Coronavirus (COVID-19) dashboard [R/OL]. (2021-07-23)[2021-07-23]: https://covid19.who.int/.

[2] BURKI T K. Lifting of COVID-19 restrictions in the UK and the Delta variant[J]. Lancet Respir Med ,2021,9(8):e85. DOI:10.1016/ S2213-2600(21)00328-3.

[3] CHEN N S, ZHOU M, DONG X, *et al*. EpidemioLogical and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study [J]. Lancet, 2020, 395(10223): 507-513. DOI: 10.1016/S0140-6736(20)30211-7.

[4] CHEN T, WU D, CHEN H,*et al*. Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study[J]. BMJ, 2020: m1295. DOI:10.1136/bmj.m1091.

[5] HUANG C L, WANG Y M, LI X W, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China[J]. Lancet, 2020, 395(10223): 497-506. DOI:10.1016/S0140-6736(20)30183-5.

[6] GUAN W J, NI Z Y, HU Y, *et al*. Clinical characteristics of coronavirus disease 2019 in China[J]. N Engl J Med, 2020, 382(18): 1708-1720. DOI:10.1056/nejmoa2002032.

[7] LI L Q, HUANG T, WANG Y Q, *et al*. COVID-19 patients' clinical characteristics, discharge rate, and fatality rate of meta-analysis[J]. J Med Virol, 2020, 92(6): 577-583. DOI:10.1002/jmv.25757.

[8] ZHOU F, YU T, DU R H, *et al*. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study [J]. Lancet, 2020, 395(10229): 1054-1062. DOI: 10.1016/S0140-6736(20)30566-3.

[9] XU F M, CHEN X, YIN X R, *et al*. Prediction of disease progression of COVID-19 based upon machine learning[J]. Int J Gen Med, 2021, 14: 1589-1598. DOI: 10.2147/IJGM.S294872.

[10] WHO. Clinical management of severe acute respiratory infection when Novel coronavirus (nCoV) infection is suspected: interim guidance [S/OL]. (2020-03-13) [2021-10-27]: https://www.who.int/docs/default-source/coronaviruse/clinical-management-of-novel-cov.

[11] JI D, ZHANG D, XU J, *et al*. Prediction for progression risk in patients with COVID-19 pneumonia: the CALL score [J]. Clin Infect Dis, 2020, 71(6): 1393-1399. DOI:10.1093/cid/ciaa414.

[12] ZHOU Y L, ZHANG Z C, TIAN J, *et al*. Risk factors associated with disease progression in a cohort of patients infected with the 2019 novel coronavirus[J]. Ann Palliat Med, 2020, 9(2): 428-436. DOI:10.21037/apm.2020.03.26.

[13] WU C M, CHEN X Y, CAI Y P, *et al*. Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China[J]. JAMA Intern Med, 2020, 180(7): 934. DOI:10.1001/jamainternmed.2020.0994.

[14] YE Z, SONG B. Liver injury in COVID-19: Diagnosis and associated factors[J]. Liver Int, 2020, 40(8): 2040-2041. DOI:10.1111/liv.14501.

[15] YIP T C F, LUI G C Y, WONG V W S, *et al*. Liver injury is independently associated with adverse clinical outcomes in patients with COVID-19[J]. Gut, 2021, 70(4): 733-742. DOI:10.1136/gutjnl-2020-321726.

[16] ZHANG Y, ZHENG L, LIU L, *et al*. Liver impairment in COVID-19 patients: a retrospective analysis of 115 cases from a single centre in Wuhan city, China[J]. Liver Int, 2020, 40(9): 2095-2103. DOI:10.1111/liv.14455.

(**收稿**:2021-07-26;**修回**:2021-08-24)

(编辑 栾 嘉)