DOI:10.12138/j.issn.1671－9638.20244826

·论著·

基于 GBDT 模型的医院室内空气微生物浓度预测

杨光飞[1,2],邬　水[3],钱翔宇[2],杨宇红[4],孙　野[5],邹　韵[6],庚俐莉[7],刘　媛[8]
(1. 大连理工大学附属中心医院,辽宁 大连　116000;2. 大连理工大学系统工程研究所,辽宁 大连　116024;3. 大连理工大学环境学院,辽宁 大连　116024;4. 大连理工大学附属肿瘤医院离退休工作部,辽宁 沈阳　110042;5. 大连理工大学附属肿瘤医院疾病预防与感染控制办公室,辽宁 沈阳　110042;6. 大连理工大学附属肿瘤医院教学与学生工作部,辽宁 沈阳　110042;7. 大连理工大学附属中心医院感染性疾病科,辽宁 大连　116000;8. 大连理工大学附属中心医院呼吸与危重症科,辽宁 大连　116000)

[摘　要] **目的**　探究基于实时室内空气环境监测数据与机器学习算法的医院室内空气微生物浓度预测。**方法**　选取 2022 年 5 月 23 日—6 月 5 日某院四个位置为监测采样点,采用物联网传感器实时监测多种空气环境数据,匹配各点位采集的空气微生物浓度数据,使用梯度提升树算法(GBDT)对医院室内空气微生物浓度进行实时预测,并选取其他五种常见的机器学习模型进行比较,对比模型包括随机森林(RF)、决策树(DT)、最近邻(KNN)、线性回归(LR)和人工神经网络(ANN)。最后通过平均绝对误差(MAE)、均方根误差(RMSE)和平均绝对百分比误差(MAPE)三个指标验证模型的有效性。**结果**　GBDT 模型在门诊电梯间(A 点)、支气管镜诊间(B 点)、CT 候诊区(C 点)和供应室护士站(D 点)的 MAPE 值分别为 22.49%、36.28%、29.34%、26.43%,GBDT 模型在三个采样点的平均性能高于其他机器学习模型,仅在一个采样点略低于 ANN 模型。GBDT 模型在四个点位的平均 MAPE 值为 28.64%,即预测值偏离实际值 28.64%,说明 GBDT 模型预测结果较好,预测值在可用范围内。**结论**　基于实时室内空气环境监测数据的 GBDT 机器学习模型能够提高医院室内空气微生物浓度预测精度。

[关　键　词] 微生物浓度;室内环境;GBDT 模型;空气微生物浓度
[中图分类号] R126.4　R197.323.4

Prediction of microbial concentration in hospital indoor air based on gradient boosting decision tree model

YANG Guang-fei[1,2], *WU Shui*[3], *QIAN Xiang-yu*[2], *YANG Yu-hong*[4], *SUN Ye*[5], *ZOU Yun*[6], *GENG Li-li*[7], *LIU Yuan*[8] (1. *Central Hospital of Dalian University of Technology, Dalian* 116000, *China*; 2. *Institute of Systems Engineering, Dalian University of Technology, Dalian* 116024, *China*; 3. *School of Environmental Science and Technology, Dalian University of Technology, Dalian* 116024, *China*; 4. *The Retired-serving Department, Cancer Hospital of Dalian University of Technology, Shenyang* 110042, *China*; 5. *Office of Disease Prevention and Infection Control, Cancer Hospital of Dalian University of Technology, Shenyang* 110042, *China*; 6. *Teaching and Student Affairs Department, Cancer Hospital of Dalian University of Technology, Shenyang* 110042, *China*; 7. *Department of Infectious Diseases, Central Hospital of Dalian University of Technology, Dalian* 116000, *China*; 8. *Department of Pulmonary and Critical Medicine, Central Hospital of Dalian University of Technology, Dalian* 116000, *China*)

[收稿日期]　2023－08－09
[基金项目]　国家自然科学基金面上项目(42071273)
[作者简介]　杨光飞(1981－),男(汉族),江苏省南京市人,教授,主要从事大数据与智能决策相关研究。
[通信作者]　杨宇红　　E-mail:1770517747@qq.com

[**Abstract**] **Objective** To explore the prediction of hospital indoor microbial concentration in air based on real-time indoor air environment monitoring data and machine learning algorithms. **Methods** Four locations in a hospital were selected as monitoring sampling points from May 23 to June 5, 2022. The "internet of things" sensor was used to monitor a variety of real-time air environment data. Air microbial concentration data collected at each point were matched, and the gradient boosting decision tree (GBDT) was used to predict real-time indoor microbial concentration in air. Five other common machine learning models were selected for comparison, including random forest (RF), decision tree (DT), k-nearest neighbor (KNN), linear regression (LR) and artificial neural network (ANN). The validity of the model was verified by the mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE). **Results** The MAPE value of GBDT model in the outpatient elevator room (point A), bronchoscopy room (point B), CT waiting area (point C), and nurses' station in the supply room (point D) were 22.49%, 36.28%, 29.34%, and 26.43%, respectively. The mean performance of the GBDT model was higher than that of other machine learning models at three sampling points and slightly lower than that of the ANN model at only one sampling point. The mean MAPE value of GBDT model at four sampling points was 28.64%, that is, the predicted value deviated from the actual value by 28.64%, indicating that GBDT model has good prediction results and the predicted value was within the available range. **Conclusion** The GBDT machine learning model based on real-time indoor air environment monitoring data can improve the prediction accuracy of indoor air microbial concentration in hospitals.

[**Key words**] microbial concentration; indoor environment; GBDT model; air microbial concentration

空气污染不仅存在于室外，室内的空气污染更值得关注[1]。随着社会模式和生活方式的改变，人们 90%的时间都在室内度过[2]，因此人们对室内空气质量的研究兴趣日益高涨。长期暴露在通风条件差或空气质量差的室内环境中会引发一系列健康问题，如室内可吸入颗粒物 $PM_{2.5}$ 或 PM_{10} 可能会引发心血管和呼吸系统等疾病[3]，挥发性有机化合物(VOCs)会对肝、肾、中枢神经系统和上呼吸道等产生毒性或致癌的作用[4]，二氧化碳(CO_2)浓度高时也可能会不同程度地影响人们的大脑认知功能[5]。除无机污染物外，室内空气中的生物气溶胶也是不可忽视的污染物。生物气溶胶是由微生物或生物衍生材料组成的悬浮生物颗粒，来自人体、空调系统、室外空气等多种污染源[6-7]，可能会导致过敏、感染传染病、真菌中毒等疾病[8-9]。当聚焦于医院这个特殊的公共场所时，生物气溶胶成为重点关注对象。研究[10]表明，空气中的微生物传播是医院感染的主要传播途径之一。已知的主要致病菌如化脓性链球菌、结核分枝杆菌和白喉棒状杆菌等，可以通过感染者的空气飞沫传播引起医院感染[11-12]，这可能会增加医院内的工作人员、探视人员，尤其是敏感群体及病患的感染风险。据估计，空气传播的细菌引起 10%～20%的医院感染[13]，因此对医院的空气微生物监测和预防是医院感染防控的工作重点。

室内空气微生物浓度可以较直观地反映感染风险，然而传统的空气微生物监测存在许多弊端。根据国家规范《医院空气净化管理规范》规定，医院只在空气消毒后采样，不能反映日常情况下真实的就医环境；规范只要求对感染高风险部门如手术部(室)、重症监护病房等的空气净化与消毒质量进行监测，对其他部门没有明确规定；如果在医院感染暴发等特殊情况出现时再进行微生物采样监测，会错过最佳的溯因时间和防控机会窗口。医院感染的预防与管理严重缺乏依据，对普通病区存在消毒与监测过度与不足并存的现象[14]。传统的空气微生物采样是一项消耗人力与物力的试验，需要专门的人员在特定采样点对空气中的微生物进行采样，然后经过培养、计数等一系列操作才能获得采样点的空气微生物浓度水平，不具备实时的特点，也就不能及时反映感染风险[15]。

当前许多研究[16-19]表明，空气中微生物浓度与 $PM_{2.5}$、CO_2、温度等环境变量存在显著相关性，但很少研究利用该相关性来实现微生物浓度预测。张铭健等[20]回顾室内微生物污染水平预测关键技术，发现一些研究使用多元线性回归的方法评估颗粒物浓度预测空气微生物浓度水平的可行性，但此类研究使用模型重复单一、相关环境变量选择有限，预测精度不能保证。本研究利用多种环境变量评估微生物(空气中的细菌)浓度的可预测性，分析空气中细菌总数，以及不同粒径范围的细菌浓度与多种环境变量、人流量之间的相关性，并使用梯度提升树(gradient boosting decision tree, GBDT)算法预测各点位的空气微生物浓度。采用物联网传感器实时监测

的环境变量评估医院空气质量及其变化规律，为医院感染防控提供直接的判断依据。

1　材料与方法

1.1　材料来源　数据来自我国东北地区某三甲医院，监测时间 2022 年 5 月 23 日—6 月 5 日，共 14 d。空气微生物监测选取该院的 4 个地点，分别是门诊电梯间（A）、支气管镜诊间（B）、CT 候诊区（C）和供应室护士站（D），监测采样地点的特征见表 1。

1.1.1　微生物监测数据　采用六级撞击式空气微生物采样器（苏州宏瑞净化科技有限公司，FSC-A6 型）对空气中浮游的微生物气溶胶采样。采样目标是细菌，培养基为大豆酪蛋白琼脂（TSA）培养基。

表 1　空气微生物监测 4 个采样点的特征

Table 1　Characteristics of the four sampling points of indoor air microbial concentration monitoring

编号	采样点	标本量（份）	消毒环境类型	潜在污染源	通风方式
A	门诊电梯间	42	Ⅳ类	人类活动（患者、医务人员、来访者）室外	中央空调 + 自然通风
B	支气管镜诊间	42	Ⅲ类	人类活动（患者、医务人员）室外	中央空调 + 自然通风
C	CT 候诊区	42	Ⅳ类	人类活动（患者、医务人员、来访者）室外	中央空调 + 自然通风
D	供应室护士站	42	Ⅲ类	人类活动（患者、医务人员、来访者）室外	中央空调 + 自然通风

在 4 个采样点，每日进行 3 次采样，每个点位共计采样 42 次，每次采样时间为 5 min，流量为 28.3 L/min，高度为 1.3 m。所有标本做好标记后在 3 h 内放入恒温培养箱，35℃培养 48 h 后取出计菌落数，计算公式如下：

$$\text{空气中细菌总数(CFU/m}^3\text{)} = \frac{\text{所有平皿菌落数(CFU)}}{\text{采样时间(min)} \times 28.3\text{(L/min)}} \times 1\,000$$

1.1.2　环境变量监测数据　环境变量数据由部署在医院的物联网传感器（上海蓝居智能科技有限公司，U-MINI208 室内环境监测终端）监测，传感器集成了几个模块，可以同时监测细颗粒物（$PM_{2.5}$）、可吸入颗粒物（PM_{10}）、CO_2、温度、相对湿度、甲醛（CH_2O）和 VOCs，传感器的测量参数细节见表 2。此外，试验人员在现场记录采样时间段内采样点的人数作为人流量数据。

1.2　机器学习模型构建

1.2.1　数据准备　如图 1 所示，本研究数据准备分两个阶段。第一阶段：按照微生物采样时间，将空气微生物数据与环境监测数据匹配。本研究中微生物采样时间为 5 min，传感器检测频率为 1 min/次，将环境监测数据均值与对应点位的空气微生物浓度数据匹配。第二阶段：利用 Prophet 时间序列预测模型[21]对缺失的环境监测数据进行填补，Prophet 算法可拟合时间序列的增长趋势、季节趋势、节假日效应及误差项获得时间序列的预测值，具有良好的预测效果。

表 2　环境监测传感器的测量参数

Table 2　Measurement parameters of the environment monitoring sensor

监测参数	缩写	范围值	分辨率	精度
温度	TEMP	－10～＋60℃	0.1℃	＜±0.5℃
相对湿度	HUMI	0～95%R.H	0.R.H	＜±5.0%R.H
细颗粒物	$PM_{2.5}$	0～1 000 μg/m³	0.1 μg/m³	＜±10%
可吸入颗粒物	PM_{10}	0～2 000 μg/m³	0.1 μg/m³	＜±10%
二氧化碳	CO_2	400～2 000 ppm	1 ppm	＜±10%
甲醛	CH_2O	0～1.5 mg/m³	0.001 mg/m³	＜±10%
挥发性有机物	VOCs	0～6 mg/m³	0.001 mg/m³	＜±10%

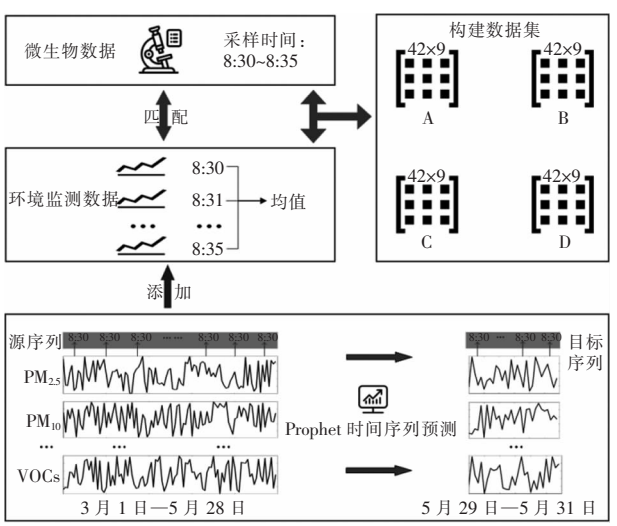


图 1　数据预处理

Figure 1　Data preprocessing

由于空气质量传感器在研究期间存在暂时的离线情况，丢失了 5 月 29—31 日 3 d 的环境监测数据，无法匹配此 3 d 内每个采样点位的 9 条细菌浓度数据。因此研究者分别提取了 3 月 1 日—5 月 28 日 3 个月内匹配采样时间的环境数据作为源序列，基于 Prophet 算法预测每个环境变量在未来 3 d 内对应采样时间的 9 个目标序列。将预测结果按照第一阶段方法与空气微生物数据匹配，构建 4 个监测点位的 4×42×9 维矩阵数据集，其中输入特征共 8 个，包括表 2 所示的 7 个环境变量和人流量变量，输出变量为空气微生物浓度。

为验证 Prophet 算法预测空气变量作为模型输入的可行性，本研究做出如下验证试验。基于 2022 年 7 月 1 日—2023 年 7 月 1 日 4 个采样点(A、B、C、D)一年的环境监测数据，提取每个采样点 9:00 共 4×365条数据进行验证。具体做法如下：为保证与源试验数据口径对齐，验证试验基于每个点位 365 d 的环境数据，使用 90 d 的数据作为源数据，使用 Prophet 算法预测未来 3 d 的目标序列，每次预测时间间隔为 30 d，共预测 26 d 的环境数据。将预测值与真实值进行对比，使用相对误差指标进行精度检验，试验结果如图 2 和表 3 所示。由于篇幅限制，图 2 仅展示了 A 点位 CO_2 预测结果。可以看出，CO_2 的预测的上下界基本涵盖真实值，且预测值与真实值保持较高一致性。各个环境变量的平均预测误差率较低，在 5.99%～34.43%，属于可用范围。见表 3。

1.2.2　GBDT 模型　GBDT 是一种基于集成学习的机器学习算法，该算法采用损失函数的负梯度作为残差近似，并通过逐渐减小残差值最小化损失函数。与现有研究[15,22-24]中常用的回归模型相比，GBDT 可以更灵活地在输入特征上实现非线性和交叉变换，以捕获空气微生物与环境变量之间不连续、非线性的关系，并且该算法无需严格的数据分布假设，对异常值具有鲁棒性和可拓展性，能够自然地对非线性决策边界进行建模[25]。GBDT 算法的学习过程和伪代码见图 3、4。(1)初始化一颗决策树来拟合输入数据。(2)在每次迭代中，计算损失函数的负梯度在当前模型的值，将其作为残差的估计；然后估计决策树叶节点区域，以拟合残差的近似值，最后利用线性搜索估计叶节点区域的值，使损失函数最小化，生成一个新的决策树。(3)通过将每一步的决策树加入到原始模型中，得到一个强学习器。

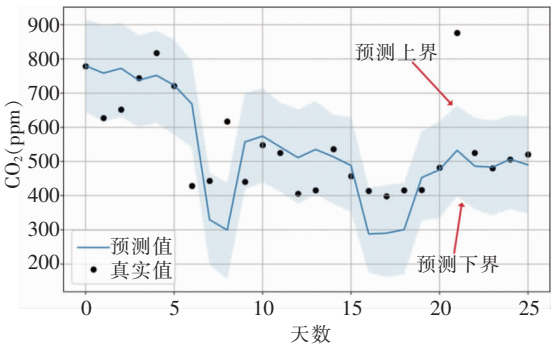


图 2　A 点位 CO_2 预测结果

Figure 2　CO_2 prediction results at point A

表 3　各点位环境变量预测值相对误差(%)

Table 3　Relative error of predicted variables of environment at each point (%)

点位	TEMP	HUMI	$PM_{2.5}$	PM_{10}	CO_2	CH_2O	VOCs
A	8.83	20.77	42.92	32.59	11.80	47.93	37.90
B	7.37	30.55	32.61	30.80	18.66	18.34	27.21
C	4.48	36.90	24.45	29.02	29.17	23.60	35.68
D	3.27	24.35	30.10	28.30	7.18	39.69	36.95
均值	5.99	28.14	32.52	30.18	16.70	32.39	34.43

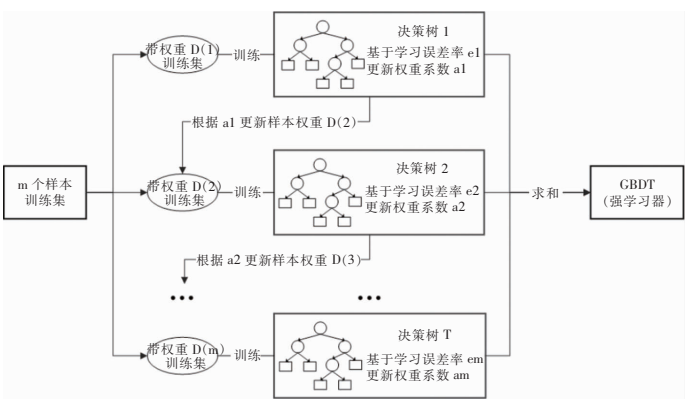


图 3　GBDT 算法的学习过程

Figure 3　Learing process of GBDT algorithm

算法：梯度提升树算法

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，$x_i\subseteq R_n$，$y_i\subseteq R$；

损失函数 $L[y,f(x)]$；

输出：回归数 $\hat{f}(x)$

1：初始化：$f_0(x)=\arg\min_c \sum_{i=1}^{N} L(y_i,c)$

2：对 $m=1,2,\cdots,M$

3：对 i=1,2,$\cdots$,N,计算 $r_{mi}=-\left[\frac{\alpha L[y_i,f(x_i)]}{\alpha f(x_i)}\right]_{f(x)=f_{m-1}(x)}$

4：对 r_{mi} 拟合一个回归树，得到第 m 棵树的叶节点区域 R_{mj}，$j=1,2,\cdots,J$

5：对 $j=1,2,\cdots,J$，计算 $c_{mj}=\arg\min_c \sum_{x_i\in R_{mj}} L[y_i,f_{m-1}(x_i)+c]$

6：更新 $f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J} c_{mj}I(x_i\in R_{mj})$

7：得到回归树 $\hat{f}(x)=f_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J} c_{mj}I(x\in R_{mj})$

图 4　GBDT 算法的伪代码

Figure 4　Pseudocode of GBDT algorithm

1.3　模型的验证　为评估所提出方法的有效性，通过测试集验证和平均绝对误差（mean absolute error，MAE）、均方根误差（root mean square error，RMSE）和平均绝对百分比误差（mean absolute percentage error，MAPE），计算评估模型的预测能力。本研究数据集涉及 4 个采样点，每个采样点的面积大小、通风情况及人流量等外界因素不同，所以空气微生物浓度高低及其分布也不相同。使用 MAE 和 RMSE 评估不同模型在同一采样点的预测误差，使用 MAPE 评估同一模型在不同采样点的预测误差。指标的计算方式如式（1）-（3）所示：

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (1)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (2)$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{y_i}\times 100\% \quad (3)$$

1.4　统计分析　应用 SPSS 25.0 软件进行统计分析。由于各采样点数据有限，不能保证数据都呈正态分布，斯皮尔曼相关系数不易受离群值影响，适用于非线性相关关系的探究，故选用斯皮尔曼相关系数探究空气微生物浓度与环境变量之间的关系。

2　结果

2.1　描述性统计

2.1.1　采样点环境变量分布特征　监测时间段内 4 个点位平均气温 25～27℃，符合国家现行的 GB/T 18883—2022《室内空气标准》中对夏季温度的要求（22～28℃），但 B 点的最高温度达 28.96℃，超出标准，原因可能是支气管镜诊间的面积小、通风不及时。相对湿度为 27.20%～50.76%，而夏季相对湿度的标准是 40%～80%，所以存在相对湿度过低的情况。室内 CO_2 最高浓度达 918.4 mg/L，平均浓度在 500～600 mg/L。采样点 VOCs 浓度最高可达 1.2 mg/m^3，CH_2O 平均浓度在 0.15 mg/m^3 左右，装修材料和清洁剂可能是主要来源。见表 4。

表 4　采样点环境变量描述性统计表

Table 4　Descriptive statistical table of environment variables at sampling points

采样点	统计指标	温度（℃）	相对湿度（%R.H）	$PM_{2.5}$（μg/m^3）	PM_{10}（μg/m^3）	CO_2（ppm）	CH_2O（mg/m^3）	VOCs（mg/m^3）	人流量（人/5 min）
A	Min	24.60	27.20	12.40	13.00	410.00	0.04	0.001	1
	Max	26.70	47.44	96.00	104.60	738.40	0.38	1.200	30
	Mean	25.62	38.72	36.71	42.15	506.96	0.14	0.640	10
	SD	0.51	4.61	19.51	22.35	94.84	0.09	0.470	8
B	Min	21.85	29.22	0.80	1.80	418.60	0.08	0.002	1
	Max	28.96	47.66	43.60	56.20	784.20	2.49	1.200	5
	Mean	25.60	39.19	17.42	20.48	485.94	0.59	0.420	3
	SD	1.61	4.37	12.24	15.40	71.57	0.52	0.440	2
C	Min	25.08	33.67	4.20	4.80	406.60	0.01	0.012	1
	Max	27.29	50.76	112.80	118.00	918.40	0.59	1.190	123
	Mean	26.50	40.58	31.73	34.40	576.08	0.15	0.610	45
	SD	0.47	4.67	24.77	26.34	160.06	0.15	0.460	38
D	Min	24.42	33.01	9.00	9.80	426.55	0.07	0.003	1
	Max	26.90	45.12	70.40	77.00	742.80	0.43	0.290	3
	Mean	26.27	37.88	35.26	41.15	541.45	0.18	0.090	1
	SD	0.45	3.03	15.70	19.14	99.65	0.09	0.060	1

注：Min 为最小值，Max 为最大值，Mean 为平均值，SD 为标准差。

各采样点的颗粒物平均浓度均符合标准(24 h 平均值 PM_{10}≤0.1 mg/m³，$PM_{2.5}$≤0.05 mg/m³)，但浓度波动很大且表现出一定的规律。见图 5。不同位点的 PM 浓度水平不同，但波动规律整体表现一致。PM 值通常会受到开窗通风和人群搅动的影响，如 PM 值在零点时处于一天中的较低水平，早晨 7:00～8:00 上班后会持续波动上升，中午左右达到最高水平，之后持续波动下降，下班后快速下降直到第二天 0:00。周末和节假日基本会出现浓度水平明显降低的情况，而节假日过后的第一天往往是医院接诊的高峰期，支气管镜诊间和 CT 候诊区可能会有特殊的就诊情况使得周末特征不明显，这都符合被监测医院的工作时间及节假日特点。A 点电梯间 PM 值的波动幅度较大，浓度忽高忽低，可能是因为电梯间的人群流动往往呈现出间歇性聚集的状态，如聚集在一起等电梯或者集中出电梯，符合实际情况。CH_2O、CO_2 与 PM 序列趋势相似，都在中午达到高峰，凌晨降至低峰。除装修情况外，白天的 CH_2O 还可能来自于消毒剂的挥发，而 CO_2 则与人的呼吸密切相关。

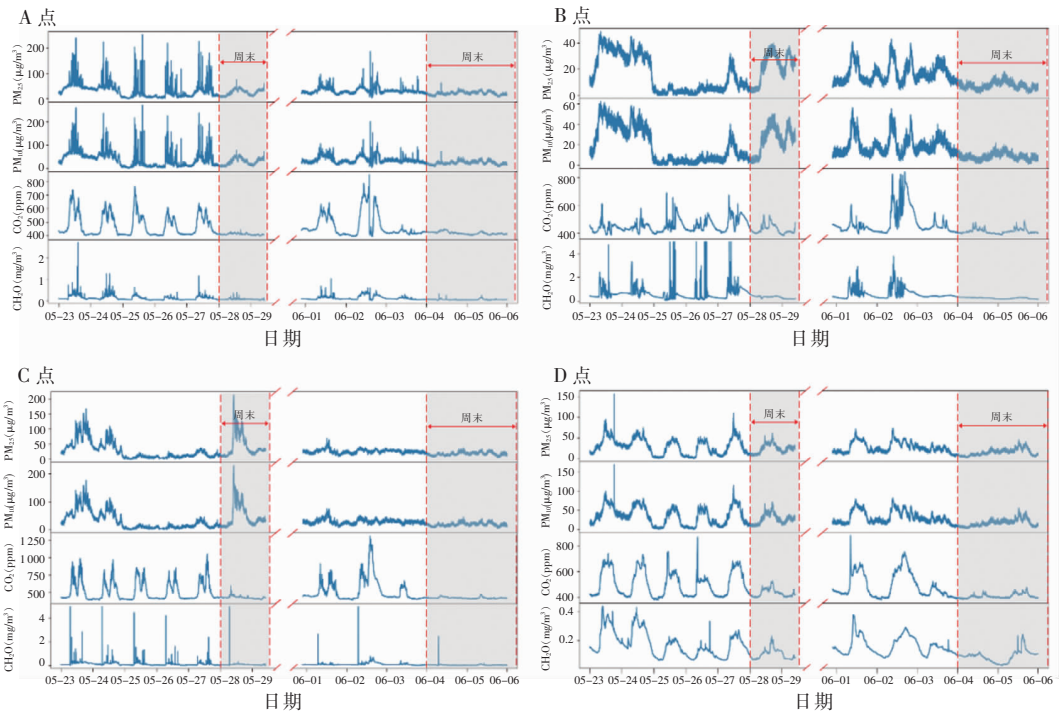


图 5　采样点环境变量时间序列趋势图

Figure 5　Time series trend of environmental variables at sampling points

2.1.2　采样点细菌浓度分布特征　监测医院 4 个点位的细菌浓度水平差异较大，14～2 466 CFU/m³，其中，A 点浓度为 49～876 CFU/m³，B 点为 21～580 CFU/m³，C 点为 42～2 466 CFU/m³，D 点为 14～1 682 CFU/m³。小提琴图见图 6。4 个采样点细菌浓度分布不相同，与采样位置的特点有密切联系。A 点和 C 点细菌浓度值分散且整体水平较高，B 点和 D 点细菌浓度值则更集中于较低水平。直观原因是门诊电梯间和 CT 候诊区属于开放型空间，人流量大，而支气管镜诊间和供应室护士站属于较封闭空间，人员少且活动范围固定。此外，C 点和 D 点存在一些极端的离群值，可能是由于候诊区患者集中就医或护士站集体开会，或是突发情况所导致的，如随患者移动的病床或大型仪器等剧烈移动。

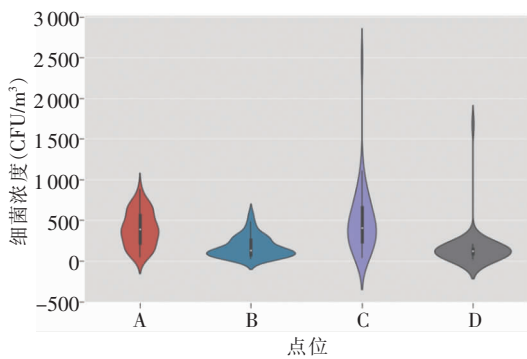


图 6　采样点细菌浓度分布图

Figure 6　Distribution of bacterial concentration at sampling points

细菌采样使用六级空气微生物撞击式采样器，模拟人呼吸道的解剖结构和空气动力学生理特征，

采用惯性撞击原理进而将悬浮在空气中的微生物粒子按照微生物粒径大小分等级地收集到无菌平皿 1 至无菌平皿 6 上，每个无菌平皿捕获粒子大小分别是第一级≥7.0 μm（皿 1）、第二级 4.7～<7.0 μm（皿 2）、第三级 3.3～<4.7 μm（皿 3）、第四级 2.1～<3.3 μm（皿 4）、第五级 1.1～<2.1 μm（皿 5）、第六级 0.65～<1.1 μm（皿 6）。皿 1～皿 6 表示第一级到第六级的细菌浓度。

人群活动量更大的 A 点和 C 点各粒径细菌浓度均高于 B 点和 D 点，其中人流量最高的 C 点各粒径细菌浓度也最高，见图 7（左）。可能是因为人群的流动带动了空气流动，导致细菌颗粒物难以沉降，并且细菌的主要来源之一就是人体，密集的人群导致了空气中的细菌浓度水平显著提升。图 7（右）反映每个点位各粒径细菌浓度的占比，A、B、D 点的皿 4、皿 5 和皿 6 所占的细菌数量均超过细菌总数的 60%，点 C 也在 50%左右，说明空气中的细菌以 3.3 μm 以下的小粒径为主。这些小粒径细菌更容易附着在 $PM_{2.5}$ 颗粒物上，通过呼吸道进入人体肺部，相较于大粒径细菌更容易带来感染风险。

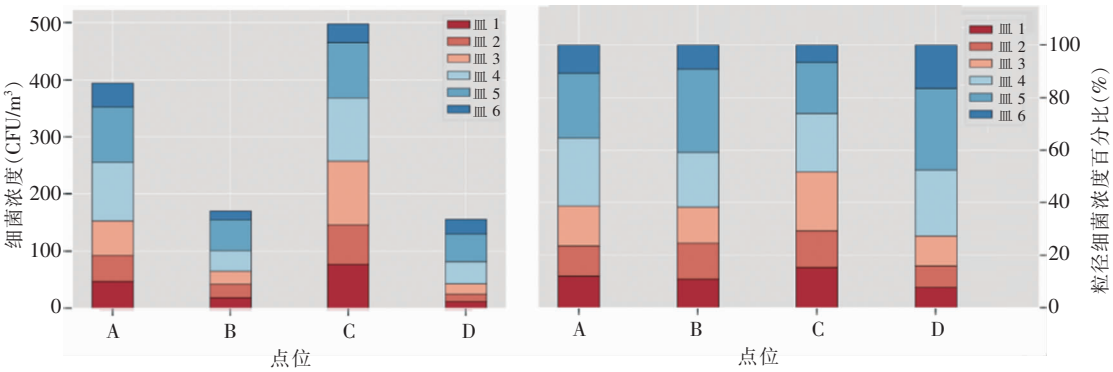


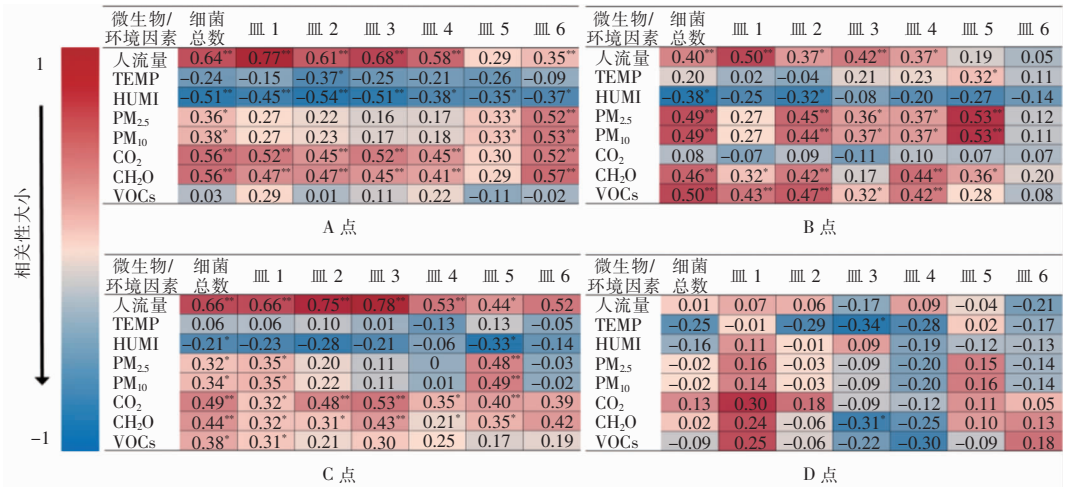
图 7　采样点六级细菌浓度及占比对比图

Figure 7　Comparison of concentration and proportion of sixth level bacteria at sampling points

2.2　相关性分析　本研究采用斯皮尔曼系数分析空气微生物浓度与各环境变量之间的相关性，见图 8。除 D 点外，其他采样点空气中的细菌浓度均与人流量呈弱到中等正相关关系。A 点细菌浓度与 CO_2、CH_2O 呈中等正相关关系，与相对湿度呈中等负相关关系，与 $PM_{2.5}$ 和 PM_{10} 呈弱但正相关关系。在 B 点，细菌浓度与 $PM_{2.5}$、PM_{10}、CH_2O 和 VOCs 呈弱到中等正相关关系，与相对湿度呈弱负相关关系。在 C 点，细菌浓度与 CO_2、CH_2O、$PM_{2.5}$、PM_{10}、和 VOCs 呈弱但正相关关系。D 点的细菌浓度与环境变量之间的相关性并不显著，原因是医院对消毒供应室的环境卫生要求非常严格，除定期打扫消毒以外，进入供应室人员必须穿工作服、戴鞋套等以保证无污染源带入，并且工作时间护士站人流量较少，部分采样期间只有采样人员在场，D 点的人流量与空气微生物浓度相关性也不显著。

各采样点皿 1～皿 6 与环境变量之间相关性分析结果显示，较大粒径的细菌与人流量相关性更强，随着粒径的缩小，细菌浓度与 $PM_{2.5}$ 和 PM_{10} 的相关性逐渐增强。尤其是粒径为 1.1～2.1 μm 的细菌浓度更为明显，也就是粒径小的微生物更多附着在粒径大小相似的颗粒物上生存。

2.3　预测模型结果　为验证 GBDT 模型在预测空气微生物浓度水平上的性能，本研究选取其他 5 种常见的机器学习模型进行比较，对比模型包括随机森林（random forest，RF）、决策树（decision tree，DT）、最近邻（k-nearest neighbor，KNN）、线性回归（linear regression，LR）和人工神经网络（artificial neural network，ANN）。模型的输入输出如 1.2.1 节所述，7 个环境变量和人流量变量作为模型的输入，空气微生物浓度作为模型的输出。每个采样点取 80%数据作为训练集，20%数据作为测试集，取 5 次五折交叉验证平均值为模型的最终得分，结果见表 5。经过计算，GBDT 模型在 A、B、C、D 4 个采样点的 MAE 均值为 100.81，RMSE 均值为 160.65，这些数值优于其他机器学习模型，即 GBDT 模型的平均性能高于其他机器学习模型，仅在 B 点略低于 ANN 模型。同时，GBDT 模型在 4 个点位的 MAPE 均值为 28.64%，即预测值偏离实际值 28.64%，说明该模型预测结果较好，预测值在可用范围内。



注：* 为 $P<0.05$，* * 为 $P<0.01$。

图 8　微生物浓度与环境变量的斯皮尔曼相关系数矩阵

Figure 8　Spearman correlation coefficient matrix between microbial concentration and environmental variables

表 5　不同采样点不同模型平均预测误差

Table 5　Mean prediction errors of different models at different sampling points

采样点	评估指标	预测模型					
		RF	DT	KNN	LR	ANN	GBDT
A	MAE	97.20	123.58	111.74	139.56	124.19	94.47
	RMSE	123.12	167.83	138.31	177.04	153.68	128.66
	MAPE(%)	24.37	29.20	29.60	31.58	40.36	22.49
B	MAE	71.56	87.48	84.97	85.89	63.15	69.01
	RMSE	101.02	126.00	116.60	124.05	97.03	99.37
	MAPE(%)	40.03	44.42	47.88	44.53	33.38	36.28
C	MAE	216.48	193.70	248.30	322.83	202.17	187.01
	RMSE	349.10	299.85	380.45	500.70	273.35	314.55
	MAPE(%)	38.71	33.02	44.75	59.27	39.77	29.34
D	MAE	136.60	77.71	105.59	197.42	69.54	52.75
	RMSE	329.99	157.58	246.45	443.03	135.15	100.01
	MAPE(%)	36.44	34.01	31.59	61.44	31.17	26.43

2.4　对照试验　为进一步验证 Prophet 时间序列预测模型的有效性和 GBDT 模型在解决非线性关系上的优越性，本研究以 2.3 节中 GBDT 模型预测结果为对照组，设计了试验组 A 和试验组 B，具体试验设置及结果见表 6。

试验组 A：由于 2022 年 5 月 29—31 日的环境监测数据丢失，导致 A、B、C、D 4 个点位中，每个点位的 9 条细菌浓度数据无法匹配到环境数据，无法构建样本。因此实验组 A 中，删除了该 9 条样本，使用剩余 33 条数据对每个点位进行建模，模型选择、模型参数及交叉验证次数与对照组保持一致。在删除了 9 条样本后，由于数据量的减少，模型无法得到充分训练，4 个采样点位的预测结果均出现了不同程度的下降，说明样本量的大小对模型的预测精度有重要作用。如 1.2.1 节结果所示，Prophet 模型的预测在可用范围之内，其带来的样本量增加可以显著提升模型的预测精度。

试验组 B：传统的线性回归模型在建模前需要过滤掉非显著线性相关的变量，以减轻模型的参数量和复杂度，获得更好的拟合结果。但 GBDT 模型不同于传统的线性回归模型，可以拟合变量间复杂的非线性关系，因此本研究未进行特征过滤，将全部

特征添加到预测模型中，以获得更好的预测结果。为进一步验证该假设，设置试验组 B，以 2.2 节结果为基准，删除每个点位不显著相关的特征，如点位 A 删除 TEMP、VOCs 两个特征。由于点位 D 的环境对卫生要求严格，经常清洁消毒，导致无环境变量与细菌浓度显著相关，因此试验组 B 中不计算点 D 的预测结果。其余试验设置与对照组保持一致。在删除了每个点位中不显著相关的特征后，各个点位的模型预测绝大多数指标呈现显著下降趋势，说明传统线性模型的特征过滤方法并不适用于本研究的 GBDT 模型，GBDT 模型可以拟合不同变量间的复杂非线性关系。

表 6　GBDT 模型与两个试验组性能比较

Table 6　Performance comparison between GBDT model and two test groups

采样点位	评估指标	GBDT 模型	试验组 A	性能提升 A(%)	试验组 B	性能提升 B(%)
A	MAE	94.47	112.02	-18.58	106.41	-12.64
	RMSE	128.66	174.17	-35.37	176.35	-37.06
	MAPE(%)	22.49	39.96	-77.68	32.97	-46.60
B	MAE	69.01	76.91	-11.45	70.65	-2.38
	RMSE	99.37	117.50	-18.24	104.99	-5.66
	MAPE(%)	36.28	37.60	-3.64	39.07	-7.69
C	MAE	187.01	173.39	7.28	189.15	-1.14
	RMSE	314.55	322.85	-2.64	310.28	1.36
	MAPE(%)	29.34	21.72	25.97	22.44	23.52
D	MAE	52.75	64.31	-21.91	-	-
	RMSE	100.01	134.37	-34.36	-	-
	MAPE(%)	26.43	31.01	-17.33	-	-

3　讨论

本研究探索利用多种环境变量评估微生物(空气中的细菌)浓度的可行性。医院环境非常复杂，各种因素如医院设计、通风系统、温度、相对湿度、各种污染物、人口密度和消毒方法等，都会影响空气中细菌的浓度[26]。分析空气中细菌总数及不同粒径范围的细菌浓度与多种环境变量(温度、相对湿度、$PM_{2.5}$、PM_{10}、CH_2O、CO_2 和 VOCs)之间的相关性，并采用 GBDT 算法预测各点位的空气微生物浓度。通过测试集验证，以及 MAE、RMSE、MAPE 评估模型的预测能力；通过物联网传感器实时监测的环境变量，评估医院空气质量及其变化规律。

在选定的采样点中，细菌浓度水平差异显著。细菌的平均浓度为 304 CFU/m^3，研究选取的采样点位按医院消毒卫生标准分类是Ⅲ类环境或Ⅳ类环境，只有供应室护士站的细菌浓度可勉强达标，其他三点均超过标准。世界卫生组织认为，空气中的细菌总数超过 700 CFU/m^3，感染风险很大，小于 500 CFU/m^3，感染风险较小。可以发现，A 点和 C 点这样的开放空间中细菌浓度超标的可能性更大，感染风险也更大，开放型空间应作为医院感染防控的关键。

当前有研究[15,22,27]探究利用颗粒物浓度评估空气中微生物水平的可能性，但也有研究[28]认为用颗粒计数代替空气微生物采样没有足够的理论支撑。我们认为，这种不一致源于不同的研究通过不同的方法探究二者之间的相关性，有时选择的方法对当前的数据分布来说是不适用的。同时以往的研究只考虑颗粒物浓度这个单一特征，其他有关影响因素的贡献被忽略，因此得出无法使用环境变量预测空气微生物浓度的结论。Seo 等[22]证明除了不同粒径的颗粒物外，还可以通过考虑温度或相对湿度等气象条件来克服现有预测模型的局限性。本研究利用 GBDT 模型并纳入多种相关环境变量的实时监测数据，提高了空气微生物浓度预测精度。

本研究选择 GBDT 模型预测空气中的微生物浓度。一方面，GBDT 模型作为一种集成学习算法，在处理数据特征之间的非线性关系上具有优势，不需要数据满足严格的假设分布，对环境数据与空气微生物数据的非线性、非正态分布特点友好。另一

方面，由于本研究收集到的空气微生物数据存在一定的离群点，而 GBDT 模型对异常值具有鲁棒性，能够自然地对非线性决策边界进行建模，进而提升模型的预测精度。最终的试验结果也验证了我们的猜想，GBDT 模型的表现整体优于其他机器学习算法。A 点（MAPE = 22.49%）、B 点（MAPE = 36.28%）、C 点（MAPE = 29.34%）和 D 点（MAPE = 26.43%）的平均预测精度为 28.64%，表现稳定且在不同点位的预测准确度差异不大。表明利用易监测的环境变量作为替代测量方法来取代基于培养方法的空气微生物浓度水平监测可行。

在不同的采样点，环境变量与微生物浓度之间的相关性存在差异，与过往研究[16-19]所得结论一致。细菌浓度与环境变量之间的关系是复杂而不确定的，因此传统线性回归方法并非最佳建模方法，本文中对比模型结果表明 GBDT 算法可以自动识别和利用变量之间的复杂关系，并得出更准确的预测结果。

基于传统培养方法的空气微生物浓度监测需要熟练的试验人员和长期的培养过程[29]，无法获得实时的空气微生物浓度水平。本研究提出的基于机器学习算法的空气微生物浓度预测模型，通过多种环境变量实时模拟医院环境的空气中微生物浓度，提供了一种快速的空气微生物浓度测量方法，节省了大量的时间、经济和人力成本，为医院感染防控提供了一种实时的反馈机制，有助于解决由医院环境带来的感染问题。

利益冲突：所有作者均声明不存在利益冲突。

[参 考 文 献]

[1] González-Martín J, Kraakman NJR, Pérez C, et al. A state-of-the-art review on indoor air pollution and strategies for indoor air pollution control[J]. Chemosphere, 2021, 262: 128376.

[2] Klepeis NE, Nelson WC, Ott WR, et al. The national human activity pattern survey (NHAPS): a resource for assessing exposure to environmental pollutants[J]. J Expo Anal Environ Epidemiol, 2001, 11(3): 231-252.

[3] Orellano P, Reynoso J, Quaranta N, et al. Short-term exposure to particulate matter (PM_{10} and $PM_{2.5}$), nitrogen dioxide (NO_2), and ozone (O_3) and all-cause and cause-specific mortality: systematic review and Meta-analysis[J]. Environ Int, 2020, 142: 105876.

[4] Tsai WT. An overview of health hazards of volatile organic compounds regulated as indoor air pollutants[J]. Rev Environ Health, 2019, 34(1): 81-89.

[5] Du BW, Tandoc MC, Mack ML, et al. Indoor CO_2 concentrations and cognitive function: a critical review[J]. Indoor Air, 2020, 30(6): 1067-1082.

[6] Pastuszka JS, Paw UKT, Lis DO, et al. Bacterial and fungal aerosol in indoor environment in Upper Silesia, Poland[J]. Atmos Environ, 2000, 34(22): 3833-3842.

[7] Hargreaves M, Parappukkaran S, Morawska L, et al. A pilot investigation into associations between indoor airborne fungal and non-biological particle concentrations in residential houses in Brisbane, Australia[J]. Sci Total Environ, 2003, 312(1-3): 89-101.

[8] Douwes J, Thorne P, Pearce N, et al. Bioaerosol health effects and exposure assessment: progress and prospects[J]. Ann Occup Hyg, 2003, 47(3): 187-200.

[9] Menetrez MY, Foarde KK, Esch RK, et al. An evaluation of indoor and outdoor biological particulate matter[J]. Atmos Environ, 2009, 43(34): 5476-5483.

[10] 秦惠，张怡，周斌，等. 医院环境致病气溶胶感染风险及其测量方法综述[J]. 暖通空调，2017，47(5)：64-71.
Qin H, Zhang Y, Zhou B, et al. Infection risk and measurement of pathogenic aerosol in hospital environment: A review[J]. Heating Ventilating & Air Conditioning, 2017, 47(5): 64-71.

[11] Pastuszka JS, Marchwinska-Wyrwal E, Wlazlo A. Bacterial aerosol in Silesian hospitals: preliminary results[J]. Pol J Environ Stud, 2005, 14(6): 883-890.

[12] Kim KY, Kim YS, Kim D. Distribution characteristics of airborne bacteria and fungi in the general hospitals of Korea[J]. Ind Health, 2010, 48(2): 236-243.

[13] Fernstrom A, Goldblatt M. Aerobiology and its role in the transmission of infectious diseases[J]. J Pathog, 2013, 2013: 493960.

[14] 姚希，巩玉秀，张宇，等. 全国医疗机构病区环境消毒现况调查与分析[J]. 中国感染控制杂志，2020，19(6)：553-558.
Yao X, Gong YX, Zhang Y, et al. Current situation of environmental disinfection in medical institutions in China[J]. Chinese Journal of Infection Control, 2020, 19(6): 553-558.

[15] Huang HL, Lee MK, Shih HW. Assessment of indoor bioaerosols in public spaces by real-time measured airborne particles[J]. Aerosol Air Qual Res, 2017, 17(9): 2276-2288.

[16] Hiwar W, King MF, Shuweihdi F, et al. What is the relationship between indoor air quality parameters and airborne microorganisms in hospital environments? A systematic review and Meta-analysis[J]. Indoor Air, 2021, 31(5): 1308-1322.

[17] Park DU, Yeom JK, Lee WJ, et al. Assessment of the levels of airborne bacteria, Gram-negative bacteria, and fungi in hospital lobbies[J]. Int J Environ Res Public Health, 2013, 10(2): 541-555.

[18] Osman ME, Ibrahim HY, Yousef FA, et al. A study on microbiological contamination on air quality in hospitals in Egypt[J]. Indoor Built Environ, 2018, 27(7): 953-968.

[19] 孙帆，钱华，叶瑾，等. 南京市校园室内空气微生物特征[J]. 中国环境科学，2019，39(12)：4982－4988.
Sun F，Qian H，Ye J，et al. Characteristics of airborne microorganisms in school classrooms in Nanjing[J]. China Environmental Science，2019，39(12)：4982－4988.

[20] 张铭健，曹国庆，冯昕. 室内微生物污染水平预测关键技术研究综述[J]. 中国环境科学，2018，38(11)：4040－4049.
Zhang MJ，Cao GQ，Feng X. Review of key technologies for forecast of indoor microbial contamination levels[J]. China Environmental Science，2018，38(11)：4040－4049.

[21] Taylor SJ，Letham B. Forecasting at scale[EB/OL]. (2017－09－27)[2023－06－30]. https://peerj.com/preprints/3190.pdf.

[22] Seo JH，Jeon HW，Choi JS，et al. Prediction model for airborne microorganisms using particle number concentration as surrogate markers in hospital environment[J]. Int J Environ Res Public Health，2020，17(19)：7237.

[23] Mousavi MS，Hadei M，Majlesi M，et al. Investigating the effect of several factors on concentrations of bioaerosols in a well-ventilated hospital environment[J]. Environ Monit Assess，2019，191(7)：407.

[24] Tseng CH，Wang HC，Xiao NY，et al. Examining the feasibility of prediction models by monitoring data and management data for bioaerosols inside office buildings[J]. Build Environ，2011，46(12)：2578－2589.

[25] Zhou F，Zhang Q，Sornette D，et al. Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices[J]. Appl Soft Comput，2019，84：105747.

[26] Mirhoseini SH，Nikaeen M，Khanahmd H，et al. Monitoring of airborne bacteria and aerosols in different wards of hospitals－particle counting usefulness in investigation of airborne bacteria[J]. Ann Agric Environ Med，2015，22(4)：670－673.

[27] Mirhoseini SH，Nikaeen M，Satoh K，et al. Assessment of airborne particles in indoor environments：applicability of particle counting for prediction of bioaerosol concentrations[J]. Aerosol Air Qual Res，2016，16(8)：1903－1910.

[28] Landrin A，Bissery A，Kac G. Monitoring air sampling in operating theatres：can particle counting replace microbiological sampling?[J]. J Hosp Infect，2005，61(1)：27－29.

[29] Tahir MA，Zhang XL，Cheng HY，et al. Klarite as a label-free SERS-based assay：a promising approach for atmospheric bioaerosol detection[J]. Analyst，2019，145(1)：277－285.

（本文编辑：左双燕）

本文引用格式：杨光飞，邹水，钱翔宇，等. 基于 GBDT 模型的医院室内空气微生物浓度预测[J]. 中国感染控制杂志，2024，23(7)：787－797. DOI：10.12138/j.issn.1671－9638.20244826.

Cite this article as：YANG Guang-fei，WU Shui，QIAN Xiang-yu，et al. Prediction of microbial concentration in hospital indoor air based on gradient boosting decision tree model[J]. Chin J Infect Control，2024，23(7)：787－797. DOI：10.12138/j.issn.1671－9638.20244826.