

DOI: 10.19663/j.issn2095-9869.20221124002

http://www.yykxjz.cn/

王印庚, 于永翔, 蔡欣欣, 张正, 王春元, 廖梅杰, 朱洪洋, 李昊. 基于 Deep Forest 算法的对虾急性肝胰腺坏死病(AHPND)预警数学模型构建. 渔业科学进展, 2024, 45(3): 171-181

WANG Y G, YU Y X, CAI X X, ZHANG Z, WANG C Y, LIAO M J, ZHU H Y, LI H. Construction of an early warning mathematical model for *Penaeus vannamei* AHPND based on the Deep Forest algorithm. Progress in Fishery Sciences, 2024, 45(3): 171-181

基于 Deep Forest 算法的对虾急性肝胰腺坏死病 (AHPND)预警数学模型构建*

王印庚^{1,2①} 于永翔^{1,2} 蔡欣欣¹ 张正^{1,2} 王春元¹
廖梅杰^{1,2} 朱洪洋¹ 李昊¹

(1. 中国水产科学研究院黄海水产研究所 山东 青岛 266071;

2. 青岛海洋科技中心海洋渔业科学与食品产出过程功能实验室 山东 青岛 266071)

摘要 为预报池塘养殖凡纳对虾(*Penaeus vannamei*)急性肝胰腺坏死病(AHPND)的发生,自2020年开始,笔者对凡纳对虾养殖区开展了连续监测工作,包括与疾病发生相关的环境理化因子、微生物因子、虾体自身健康状况等18个候选预警因子指标,通过数据标准化处理后分析病原、宿主与环境之间的相关性,对候选预警因子进行筛选,基于Python语言编程结合Deep Forest、LightGBM、XGBoost算法进行数据建模和预测性能评判,仿真环境为Python2.7,以预警因子指标作为输入样本(即警兆),以对虾是否发病指标作为输出结果(即警情),根据输入样本和输出结果各自建立输入数据矩阵和目标数据矩阵,利用原始数据矩阵对输入样本进行初始化,结合函数方程进行拟合,拟合的源代码能利用已知环境、病原及对虾免疫指标数据对目标警情进行预测。最终建立了基于Deep Forest算法的虾体(肝胰腺内)细菌总数、虾体弧菌(*Vibrio*)占比、水体细菌总数和盐度的4维向量预警预报模型,准确率达89.00%。本研究将人工智能算法应用到对虾AHPND发生的预测预报,相关研究结果为对虾AHPND疾病预警预报建立了预警数学模型,并为对虾健康养殖和疾病防控提供了技术支撑和有力保障。

关键词 对虾;急性肝胰腺坏死病;预警数学模型;Deep Forest算法;Python语言

中图分类号 S917.1 **文献标识码** A **文章编号** 2095-9869(2024)03-0171-11

对虾急性肝胰腺坏死病(acute hepatopancreatic necrosis disease, AHPND)流行面广、致病性强、死亡率高,在全球对虾养殖地区肆虐,一旦暴发难以治疗,造成大批对虾发病死亡,给中国乃至全球对虾养殖业带来了数百亿元的经济损失(贾丹等, 2018; Soto-Rodriguez et al, 2015; 韩琳等, 2018)。从AHPND流行特点和防

控实践来看,该病传播速度快、治愈难度大,一直是国内对虾养殖业面临的严峻考验(陈蒙蒙等, 2018)。随着AHPND病理病因的深入研究,人们逐步认识到及时预警和预防能够有效地规避或减少因该疾病发生导致的经济损失,并有利于防控操作和健康养殖生产。

* 国家重点研发计划(2019YFD0900102)、山东省泰山产业领军人才项目(LJNY201802)和苏北科技专项(SZ-LYG202028)共同资助。

① 通信作者:王印庚,研究员, E-mail: wangyg@ysfri.ac.cn

收稿日期: 2022-11-24, 收修改稿日期: 2023-03-21

病害预测以病原生态学为基础,定量调查与动植物病害发生发展和传播密切相关的影响因素,运用数理分析方法,依据若干关键的主效因素建立疾病流行发展的数学模型,进而实现预报疾病的发生。符合客观规律的疾病预测的准确度和精确度在很大程度上取决于所选择的数学分类方法。目前常用的有近 40 种之多(高争光, 2020; Ha *et al.*, 2017)。近 30 年来,人工智能被广泛地应用到许多学科领域,取得的成果也较为丰富多样(Underwood *et al.*, 2021; 万强等, 2016; Hamer *et al.*, 2020)。

预警预报是目前病害防控研究的热点和焦点。近年来,国内外研究人员在农作物灾害风险预警方面取得了丰硕的研究成果,并将该技术广泛应用于农作物生产实践中的病害防控领域,对于农业病害的防控和维持农业经济稳定有较强的指示作用(Huang *et al.*, 2019; Rumpf *et al.*, 2010; 毕春光等, 2021)。我国水产动物病害预警预报的相关研究比较薄弱,主要是依据采集分析养殖水质来对鱼类疾病发生趋势进行预测预报研究(马真, 2013; 徐云娟, 2017)。水产动物疾病的发生是病原、宿主与环境因素相互作用的结果,确定与疾病发生相关的主效因子,再结合合适的数学分类算法进行疾病的预报,能保证预警模型的精确性和可靠性(Beaz-Hidalgo *et al.*, 2010; Paillard *et al.*, 2004)。而关于对虾 AHPND 预测方面的预报模型尚未有报道,研究建立对虾 AHPND 预警预报模型,不仅能为

对虾疾病的预警预报体系构建提供人工智能算法,也可为对虾养殖业的健康发展提供技术支撑。

本研究通过采集关于对虾 AHPND 发生相关的环境理化因子、微生物因子和机体自身免疫因子等 18 个候选预警因子参数,通过归一化处理 and Pearson 相关性分析,逐一剔除不相关因子,基于 Python 语言编程结合 Deep Forest 算法进行数据建模,仿真环境为 Python2.7,同时采用不同的基于决策树的优化方法进行预测准确率和性能比较,初步构建基于 Deep Forest 算法的池塘养殖凡纳对虾(*Penaeus vannamei*)急性肝胰腺坏死病的预警模型。

1 数据采集

从 2020 年 7 月开始,对于河北省沧州市某对虾养殖区域内 4 口池塘的养殖参数、发病情况进行监测。池塘面积为 2.4~3.3 hm²,投放密度约为 22.5 万尾/hm²,检测起始期凡纳对虾平均体重约为(7.0±1.0) g/只;每次采样记录凡纳对虾的体色、活力、摄食强度、健康程度和生产管理情况以及理化因子和生物因子。根据“环境-病原-宿主”三者间的辩证关系,结合 AHPND 发生相关因子的定性分析,本研究选取 18 个可能影响对虾 AHPND 发生的环境、病原及免疫因子进行监测(表 1),包括环境因子(气温、水温、盐度、pH、DO、氨氮、亚硝态氮),病原因子[水体内可培养细菌和弧

表 1 数据分析所用参数列表

Tab.1 List of parameters used in data analysis

参数 Parameters	参数代号 Parameter code	单位 Unit
气温 Air temperature	1	℃
水温 Water temperature	2	℃
盐度 Salinity	3	
酸碱度 pH	4	
溶解氧(DO) Dissolved oxygen	5	mg/L
氨氮(NH ₄ ⁺ -N) Ammonia nitrogen	6	mg/L
亚硝态氮(NO ₂ -N) Nitrite	7	mg/L
水体细菌总数 Bacteria in water	8	CFU/mL
水体弧菌总数 <i>Vibrio</i> in water	9	CFU/mL
水体弧菌占比 Proportion of <i>Vibrio</i> in water	10	%
虾体细菌总数 Bacteria in shrimp	11	CFU/g
虾体弧菌总数 <i>Vibrio</i> in shrimp	12	CFU/g
虾体弧菌占比 Proportion of <i>Vibrio</i> in shrimp	13	%
碱性磷酸酶 AKP	14	U/mg
酸性磷酸酶 ACP	15	U/mg
超氧化物歧化酶 SOD	16	U/mg
溶菌酶 LZM	17	U/mg
酚氧化酶 PO	18	U/mg

菌(*Vibrio*)总数及弧菌占比、虾体内(肝胰腺内)可培养细菌和弧菌总数及弧菌占比], 宿主免疫因子[碱性磷酸酶(AKP)、酸性磷酸酶(ACP)、溶菌酶(LZM)、超氧化物歧化酶(SOD)、酚氧化酶(PO)], 相关样品采集方法及数据信息参考于永翔等(2023)。

气温根据采样时的天气预报显示值记录, 水温、盐度、DO 和 pH 采用 YSI 便携式水质测定仪(美国)测定, 氨氮和亚硝态氮采用 WTW 6600 分光光度仪(德国)结合氨氮和亚硝态氮检测试剂盒测定, 可培养细菌和弧菌数量根据 TSB 和 TCBS 平板培养计数推算测定, 虾体和水体弧菌占比通过计算获得, 对虾肌肉免疫酶活性使用南京建成生物技术有限公司的免疫酶活试剂盒测定(于永翔等, 2023)。

2 数据预处理

2.1 剔除无效数据

同一模式条件下共采集到相关数据 140 组, 对养殖过程中采集到的额外泼洒外源投入品的组别进行剔除, 保证数据具有一定的自然客观性和科学性。通过删除无效数据, 已分类监测数据有 100 组, 包括发病数据 25 组, 健康数据 75 组。

2.2 数据标准化处理

由于不同影响因子之间存在的量纲和数量大小差异会对构建模型造成影响, 为提高后续实验训练的速度和预测的精度, 对处理好的 100 组训练测试数据采用 mapminmax 函数对其进行归一化处理, 使处理后的数据分布在 [0, 1] 之间。用 X 表示第 1 个变量, 用 Y 表示第 2 个变量, 它们分别是 $X=(x_1, x_2, \dots, x_i)$; $Y=(y_1, y_2, \dots, y_i)$, 其中, $\max(x)$ 和 $\min(x)$ 分别表示 X 变量的最大值和最小值。函数公式如下:

$$y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

2.3 预报因子的删选

对虾病害发生与环境因子、微生物因子以及对虾机体自身免疫因子密切相关, 而这些因子之间又存在错综复杂的联系, 其中一个变量的变动会引起另一些变量的联动。使用 SPSS 21 软件进行 Pearson 相关分析, 分析各个影响因子与疾病发生情况以及相互之间的相关性, 并进行假设检验计算 P 值。如果用 X 表示第 1 个变量, 用 Y 表示第 2 个变量, 它们分别是 $X=(x_1, x_2, \dots, x_n)$; $Y=(y_1, y_2, \dots, y_n)$, 那么 Pearson 相关系数由下式定义:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \sum_{i=1}^n (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2)$$

式中, \bar{x} 和 \bar{y} 分别是 2 个变量的平均数。各参数因子与预测对象的相关系数可以反映每个参数对预测对象的影响程度, 因此, 相关系数可作为筛选预测因子的依据。根据通常规定的临界值, 选取那些与预测对象的相关系数大于临界值的因子, 即将与预测对象相关性程度较高的一些因子作为预报的预选指标。

3 Deep Forest 模型构建

Deep Forest 算法的预测过程分为 2 个阶段: 多粒度扫描阶段(multi-grained scanning)和级联森林阶段(cascade forest)。Deep Forest 算法流程具体实现步骤主要是对原始数据序列进行算法的预处理, 利用 Deep Forest 算法对训练样本集进行训练, 最后利用预测样本的特征序列进行预测(林鹏程等, 2019)。

4 结果与分析

4.1 确定训练集和测试集

通过预处理的 100 组数据, 随机挑选 70% 的样本数据作为训练集, 另外的 30% 样本数据作为测试集。

4.2 对虾发病与参数因子的相关性分析

Pearson 相关分析表明(表 2), 对虾发病与虾体细菌总数、虾体弧菌总数、LZM、虾体弧菌占比、水体细菌总数、盐度、ACP、水体弧菌和 PO 等 9 个参数具有显著正相关性($P < 0.05$), 相关系数分别为 0.574 1、0.521 1、0.438 2、0.391 1、0.374 7、0.350 1、0.257 1、0.238 3 和 0.228 9, 表明在一定范围内这些参数值升高, 该病害更易发生、更严重; 与高温期水温呈显著负相关性($P < 0.05$), 其相关系数为 -0.227 9, 表明在高温期, 水温降低病害更容易发生。而气温、pH、DO、氨氮、亚硝态氮、水体弧菌占比、AKP 和 SOD 等 8 个参数与对虾发病相关性不大($P > 0.05$)。

一般界定变量相关系数的绝对值处于 0.8~1.0 之间为极强相关关系, 0.6~0.8 之间为强相关关系, 0.4~0.6 为中等程度相关关系, 0.2~0.4 为弱相关关系, 0~0.2 为极弱相关或不相关(Schober *et al*, 2018)。其中, 病害发生与虾体细菌总数、虾体弧菌总数和 LZM 的相关系数的绝对值均介于 0.4~0.6 之间, 为中等程度相关关系; 虾体弧菌占比、水体细菌总数、盐度、

表2 各参数因子与病害发生的相关性分析结果
Tab.2 Correlation analysis results of each parameter factor and disease occurrence

参数 Parameters	Pearson 相关系数 Pearson correlation	P 值检验 P value
气温 Air temperature	-0.189 6	0.059
水温 Water temperature	-0.227 9*	0.023
盐度 Salinity	0.350 1**	0.000
酸碱度 pH	0.017 9	0.859
溶解氧 (DO) Dissolved oxygen	-0.024 8	0.806
氨氮 (NH ₄ ⁺ -N) Ammonia nitrogen	0.128 2	0.204
亚硝态氮 (NO ₂ -N) Nitrite nitrogen	0.117 7	0.244
水体细菌总数 Bacteria in water	0.374 7**	0.000
水体弧菌总数 <i>Vibrio</i> in water	0.238 3*	0.017
水体弧菌占比 Proportion of <i>Vibrio</i> in water	0.037 3	0.713
虾体细菌总数 Bacteria in shrimp	0.574 1**	0.000
虾体弧菌总数 <i>Vibrio</i> in shrimp	0.521 1**	0.000
虾体弧菌占比 Proportion of <i>Vibrio</i> in shrimp	0.391 1**	0.000
碱性磷酸酶 AKP	0.010 7	0.916
酸性磷酸酶 ACP	0.257 1*	0.010
超氧化物歧化酶 SOD	-0.101 6	0.315
溶菌酶 LZM	0.438 2**	0.000
酚氧化酶 PO	0.228 9*	0.022

注: *表示相关性呈显著水平, **表示相关性呈极显著水平。下同。

Note: * indicates a significant level of correlation, and ** indicates a highly significant level of correlation. The same below.

ACP、水体弧菌、PO 和水温等参数的相关系数绝对值均介于 0.2~0.4 之间, 为弱相关性关系, 说明各种影响发病显著的参数因子影响程度也存在差异, 进一步说明各参数与病害发生的关系较为复杂。

4.3 预报因子的删选

相关性分析表明(图 1), 虾体细菌总数、虾体弧菌总数、LZM、虾体弧菌占比、水体细菌总数、盐度、ACP、水体弧菌、PO 和水温等 10 个参数与对虾 AHPND 发生都存在显著相关性, 是最相关的前 10 个特征值, 可以作为预报模型构建的预选指标。

考虑到这些参数因子之间可能相互影响, 即一个参数的变动会引起另一些参数的变动, 因此, 通过 Pearson 分析进一步探究各因子之间的相关性关系, 逐一剔除对因变量影响不显著或间接作用的因子, 因子间相关系数见表 3。

(1)监测高温期水温与虾体细菌、虾体弧菌、虾体弧菌占比、水体细菌和 LZM 等因子呈一定的负相关性关系, 且相关系数绝对值均介于 0.2~0.4 之间, 为弱相关性关系, 说明随着高温期水温降低这些因子会呈一定程度的升高。与盐度、水体弧菌、PO 和 ACP

等因子相关系数绝对值均低于 0.2, 说明监测期内水温与这些因子之间为极弱相关关系。

(2)盐度与虾体弧菌占比、PO、ACP 和 LZM 等因子呈一定的正相关性关系, 且与 PO、ACP 和 LZM 相关系数的绝对值均大于 0.4, 为中等程度相关, 与虾体弧菌占比相关系数的绝对值为 0.37, 呈弱相关, 说明盐度升高会引起这些因子不同程度的升高; 与其他因子相关系数绝对值均低于 0.2, 为极弱相关。

(3)虾体细菌与虾体弧菌、虾体弧菌占比、水体细菌、水体弧菌、PO、ACP 和 LZM 均呈正相关性关系, 因为虾体弧菌数量也包含在细菌总数中, 所以虾体弧菌直接影响虾体细菌的数量, 其相关系数绝对值为 0.91, 为极强相关, 与水体细菌、虾体弧菌占比和 LZM 相关系数绝对值介于 0.4~0.6 之间, 为中等程度相关, 与水体弧菌和 ACP 相关系数绝对值介于 0.2~0.4 之间, 为弱相关。与 PO 相关系数绝对值为 0.17, 为极弱相关。

(4)虾体弧菌与虾体弧菌占比、水体细菌、水体弧菌、PO、ACP 和 LZM 均呈正相关性关系, 其中虾体弧菌占比是虾体弧菌与细菌的比值, 受虾体弧菌数量的直接影响, 相关系数绝对值为 0.69, 为强相关;

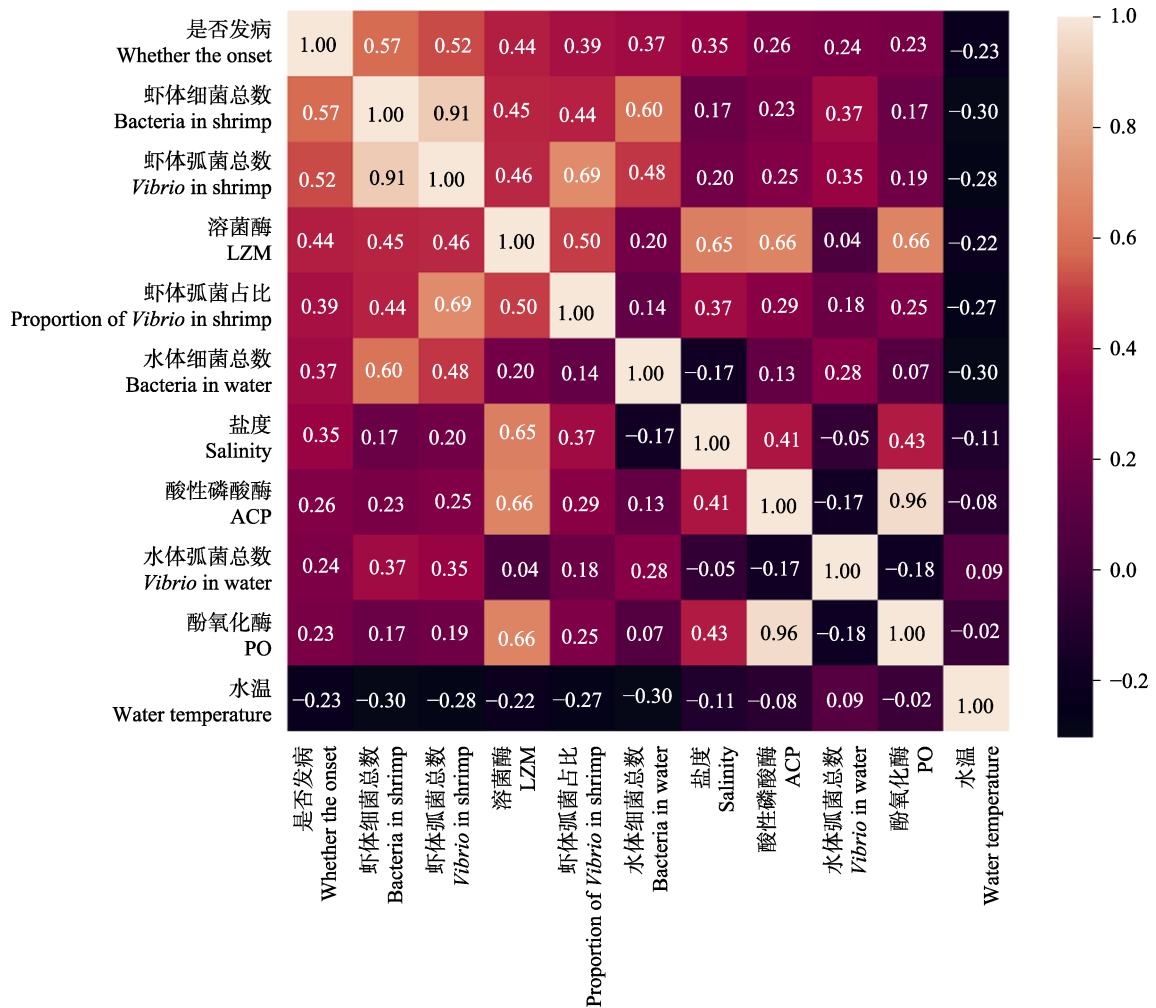


图 1 对虾 AHPND 发生最相关的 10 个特征值相关性热力图

Fig.1 Heatmap of the correlation between the 10 most relevant eigenvalues for the incidence of AHPND

与水体细菌和 LZM 相关系数绝对值介于 0.4~0.6 之间, 为中等程度相关; 与水体弧菌和 ACP 相关系数绝对值介于 0.2~0.4 之间, 为弱相关; 与 PO 相关系数绝对值为 0.19, 为极弱相关。

(5) 虾体弧菌占比与水体细菌、水体弧菌、PO、ACP 和 LZM 均呈正相关性关系, 与 LZM 相关系数绝对值为 0.49, 为中等程度相关; 与 PO 和 ACP 相关系数绝对值介于 0.2~0.4 之间, 为弱相关; 与水体细菌和水体弧菌相关系数绝对值介于 0~0.2 之间, 为极弱相关。

(6) 水体细菌与水体弧菌、PO、ACP 和 LZM 均呈一定程度的正相关性关系, 其中, 与水体弧菌和 LZM 相关系数绝对值介于 0.2~0.4 之间, 为弱相关; 与 PO 和 ACP 相关系数绝对值介于 0~0.2 之间, 为极弱相关。

(7) 水体弧菌与 PO、ACP 和 LZM 相关系数的绝对值均介于 0~0.2 之间, 为极弱相关;

(8) PO 与 ACP、LZM 相关系数的绝对值分别为

0.96 和 0.66, 均为强相关; 与 ACP 和 LZM 相关系数的绝对值为 0.66, 为强相关。

综上所述, 在与对虾发病特征排序前 10 的因子中, 虾体细菌与虾体弧菌高度相关, 与虾体弧菌占比、水体细菌和 LZM 中等程度相关, 虾体弧菌占比直接受虾体弧菌和虾体细菌的共同调控, 水体细菌与虾体弧菌也呈中等程度相关, 说明在对虾整个养殖系统中, 可培养微生物在水体和动物机体中的关联度较高; LZM 与盐度、PO 和 ACP 强相关, 与虾体细菌和虾体弧菌等中等程度相关, 盐度与 PO 和 ACP 中等程度相关, 说明环境因子和病原微生物的外界刺激可能会激发宿主免疫应激反应, 且 LZM、PO 和 ACP 可能存在一定的协同作用。在参数进行降维处理构建模型的过程中, 根据参数因子之间的相关性关系对参数进行剔除, 优先可以采用虾体细菌总数、虾体弧菌总数或虾体弧菌占比、LZM、水体细菌总数、盐度和水体弧菌作为预报模型构建的预选指标。

表3 参数因子之间的相关性分析结果
Tab.3 Correlation analysis results between parameter factors

参数 Parameters	水温 Water temperature	盐度 Salinity	虾体细菌 Bacteria in shrimp	虾体弧菌 Vibrio in shrimp	虾体弧菌占比 Proportion of Vibrio in shrimp	水体细菌 Bacteria in water	水体弧菌 Vibrio in water	酚氧化酶 PO	酸性磷酸酶 ACP	溶菌酶 LZM
水温 Water temperature	1.000 0									
盐度 Salinity	-0.111 0	1.000 0								
虾体细菌 Bacteria in shrimp	-0.296 0*	0.169 5	1.000 0							
虾体弧菌 Vibrio in shrimp	-0.281 6*	0.197 5	0.910 0**	1.000 0						
虾体弧菌占比 Proportion of Vibrio in shrimp	-0.272 1*	0.3712**	0.444 9**	0.693 5**	1.000 0					
水体细菌 Bacteria in water	-0.304 0*	-0.1712	0.604 0**	0.483 9**	0.142 6	1.000 0				
水体弧菌 Vibrio in water	0.088 2	-0.052 1	0.370 9**	0.353 1**	0.178 3	0.279 4*	1.000 0			
酚氧化酶 PO	-0.021 5*	0.431 8**	0.168 6	0.185 6	0.248 4*	0.074 0	-0.182 0	1.000 0		
酸性磷酸酶 ACP	-0.082 2	0.410 7**	0.229 5*	0.252 7*	0.291 1*	0.129 7	-0.168 2	0.961 0**	1.000 0	
溶菌酶 LZM	-0.216 6	0.649 4	0.453 7**	0.456 9**	0.496 2**	0.201 0*	0.037 5	0.663 8**	0.661 3**	1.000 0

4.4 预报模型的构建

利用 Deep Forest 算法进行模型构建的同时, 选取了 LightGBM (LGB)和 XGBoost (XGB)等 2 种基于决策树的集成学习方法算法对相同预报因子参数构建模型, 并将预测效果与 Deep Forest 预测效果进行比较分析, 评判不同维度参数和不同算法构建模型的预测效果和可行性, 进而选择最适于对虾 AHPND 发生的预报模型。本章预测算法程序均基于 Python 语言编程, 主要程序运行代码如下:

Deep Forest 算法:

```
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.model_selection import KFold
from deepforest import CascadeForestClassifier
import pandas as pd
```

LightGBM 算法:

```
import lightgbm as lgb
import pandas as pd
from sklearn.metrics import mean_squared_error
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import KFold
```

XGBoost 算法:

```
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
import xgboost as xgb
import pandas as pd
from sklearn.model_selection import KFold
```

4.5 预报模型的准确度评判

将与发病影响显著的特征排序前 10 的因子, 即: 虾体细菌总数、虾体弧菌总数、LZM、虾体弧菌占比、水体细菌总数、盐度、ACP、水体弧菌、PO 和水温, 按照相关性从大到小排列, 从低维度到高维度依次进行组合, 建立 1~10 参数的对虾 AHPND 预测模型, 并通过模型预测准确率来评判。

模型预测结果如表 4 所示, 不同维度参数组合基于不同算法构建的模型的预测准确率不同, 最高可达 89.00%。基于 Deep Forest 算法, 参数维度由 1~8 的模型准确率随着维度增加逐步提高, 同时, 随着维度增加, 其准确率降低但幅度较小, 其中, 只有 8 维模型的预测准确率最高(89.00%), 其次是 6 维和 9 维模型(88.00%), 5 维和 10 维模型准确率为 87.00%, 2~4 维模型准确率为 85.00%, 1 维模型准确率最低(73.00%), 说明维度过高或过低参数组合的模型都会影响模型的准确率。

表 4 不同维度参数组合、不同算法分类结果比较

Tab.4 Comparison of classification results of different dimension parameter combinations and different algorithms

参数个数 Number of parameters	参数组合 Combination of parameters	算法 Algorithms		
		Deep Forest	XGBoost	LightLGB
1	X1	73.00%	83.00%	75.00%
2	X1+X2	85.00%	85.00%	75.00%
3	X1+X2+X3	85.00%	83.00%	75.00%
4	X1+X2+X3+X4	85.00%	83.00%	75.00%
5	X1+X2+X3+X4+X5	87.00%	89.00%	75.00%
6	X1+X2+X3+X4+X5+X6	88.00%	85.00%	75.00%
7	X1+X2+X3+X4+X5+X6+X7	86.00%	83.00%	75.00%
8	X1+X2+X3+X4+X5+X6+X7+X8	89.00%	82.00%	75.00%
9	X1+X2+X3+X4+X5+X6+X7+X8+X9	88.00%	83.00%	75.00%
10	X1+X2+X3+X4+X5+X6+X7+X8+X9+X10	87.00%	78.00%	75.00%

注: X1~X10 分别代表虾体细菌、虾体弧菌、LZM、虾体弧菌占比、水体细菌、盐度、ACP、水体弧菌、PO 和水温。下同。

Note: X1-X10 represent bacteria in shrimp, *Vibrio* in shrimp, LZM, proportion of *Vibrio* in shrimp, bacteria in water, salinity, ACP, *Vibrio* in water, PO, and water temperature, respectively. The same below.

基于 XGBoost 算法, 参数维度 1~10 的模型准确率范围随着维度增加存在较大幅度波动, 预测模型的稳定性还需进一步研究, 只有 5 维模型准确率最高 (89.00%), 其次是 6 维模型 (85.00%)。1 维模型比基于 Deep Forest 算法的模型准确率高出 10.00%, 为 83.00%, 10 维模型比基于 Deep Forest 算法的模型准确率低 9.00%, 为 78.00%, 其精确度和可适用性还需要后续进一步验证。基于 LightGBM 算法, 参数维度 1~10 的模型准确率不随维度的变化而变化, 一直保持在 75.00%, 预测效果稳定、准确率较低, 缺乏针对性和特异性, 推测该算法可能不适用本研究数据的模型构建。3 种算法对比结果显示, Deep Forest 算法针对对虾 AHPND 发生构建的模型性能要优于 XGBoost 和 LightGBM 两种算法。

4.6 模型的进一步构建

根据参数因子之间的相关性分析结果进一步筛选剔除后, 将与发病影响显著的特征排序前 7 的因子, 即: 虾体细菌总数、虾体弧菌总数、LZM、虾体弧菌占比、水体细菌总数、盐度和水体弧菌, 按照相关性从大到小排列, 从低维度到高维度依次进行组合, 建立 2~7 维参数的对虾 AHPND 预测模型, 并通过模型预测准确率来评判。

模型预测结果如表 5 所示, 不同维度参数组合基于不同算法构建模型的预测准确率也各有不同。其中基于 Deep Forest 算法, 整体比参数因子筛选之前的同维度模型的准确率有了一定程度的提高。准确率最高 (89.00%) 的模型为 4、5、7 维参数组合, 分别比之

前提高了 4.00%、2.00%、3.00%; 其次为 88.00% 的 4 维 (序号 4) 和 6 维模型 (序号 8), 4 维模型比之前提高了 3.00%, 6 维模型准确率与之前相同; 5 维模型的准确率保持不变 (87.00%) (序号 6); 3 维和 4 维模型的准确率均提高了 1.00%, 为 86.00% (序号 2、3)。

基于 XGBoost 算法, 模型准确率随着参数维度的增加依然存在较小范围内的波动, 整体比基于 Deep Forest 算法模型的准确率偏低。其中, 准确率最高 (89.00%) 的模型依然为 4 维参数组合模型 (序号 5), 相比之前提高了 6.00%; 其次为准确率为 86.00% 的 4 维模型 (序号 3), 准确率提高了 3.00%; 准确率为 85.00% 的 4 维模型比之前提高了 2% (序号 4)、5 维模型比之前降低了 4.00% (序号 6、7), 预测效果不稳定。基于 LightGBM 算法, 所有参数维度的模型准确率一直保持在 75.00% 不变, 且远远低于另外 2 种算法, 可能对于对虾 AHPND 模型预警没有指导意义。

综合看来, Deep Forest、XGBoost、LightGBM 三种算法对于不同维度参数组合构建模型的预测效果各不相同, 其中, 基于 Deep Forest 算法的模型准确率较高, 随着参数组合维度变化, 构建的模型准确率变化也较稳定, 在一定程度上认为对于对虾 AHPND 发生具有积极的指导意义, 所以本研究优先选择 Deep Forest 算法来构建对虾 AHPND 发生的预警模型。

其中, 基于 Deep Forest 算法准确率最高 (89.00%) 的有 4、5、7 维模型, 因为高维度的预测模型参数数量较多, 较为复杂, 在实际操作过程中的工作量较大, 采集数据难以及时完成, 因此, 基于实践应用优先选择低维度、准确率高的参数组合来建立对虾 AHPND

表5 筛选因子的不同维度组合、不同算法分类结果比较

Tab.5 Comparison of classification results of different dimension combinations and different algorithms of screening factors

序号 Number	参数个数 Number of parameters	参数组合 Combination of parameters	算法 Algorithms		
			Deep Forest	XGBoost	LightGBM
1	2	X1+X4	81.00%	82.00%	75.00%
2	3	X1+X4+X5	86.00%	82.00%	75.00%
3	4	X1+X2+X4+X8	86.00%	86.00%	75.00%
4	4	X1+X2+X5+X6	88.00%	85.00%	75.00%
5	4	X1+X4+X5+X6	89.00%	89.00%	75.00%
6	5	X1+X2+X4+X5+X6	87.00%	85.00%	75.00%
7	5	X1+X4+X5+X6+X8	89.00%	85.00%	75.00%
8	6	X1+X2+X4+X5+X6+X8	88.00%	82.00%	75.00%
9	7	X1+X2+X3+X4+X5+X6+X8	89.00%	84.00%	75.00%

的预测模型。本研究选择由 X1、X4、X5 和 X6 参数所构建的 4 维模型,即虾体细菌总数、虾体弧菌占比、水体细菌和盐度 4 个参数来构建对虾 AHPND 发生预警预测模型。

5 讨论

Deep Forest 的深度学习将传统机器学习模型中的随机森林作为基础模型,是基于集成模型的深度模型,它也可以在除了大规模图像数据之外的任务中获得和深度神经网络相似的表现,通过在现有模型基础上改进能够比较快进入应用。目前,在保险购买预测场景领域(林鹏程等, 2019)、电力系统短期负荷预测领域(陈吕鹏等, 2018)已实现广泛应用。本研究构建了基于 Deep Forest 算法的 4 维预警模型即虾体细菌总数、虾体弧菌占比、水体细菌和盐度 4 个参数来对对虾 AHPND 发生进行预测预报,准确率为 89.00%。Deep Forest 算法构建的模型准确率高、预测效果稳定,对于预测对虾 AHPND 发生具有积极的实践性应用价值。其中,虾体内病原微生物含量扩增是影响对虾 AHPND 暴发的主要因素,与水体中病原微生物存在一定相互影响,和盐度共同影响导致机体的免疫力下降。有研究表明,随着盐度等环境因子的变化,引起水生动物因渗透压不适而抗病力下降,病原微生物大量扩增会在短时间内引起发病和死亡,这与本研究结果一致(Schofield *et al*, 2021)。盐度高也代表着池水有机物大量积累、病原微生物增殖的状态,意味着疾病发生的风险增大。因此,本研究在预警模型中保留盐度参数具有重要含义和意义。温度变化会导致水质变化加快,水质的不稳定容易引起疾病暴发(Karvonen *et al*, 2010; 张健龙, 2016)。池塘养殖过程中,水温受气温和降雨的影响,其改变会引起水体中

DO、pH、氨氮和亚硝态氮的联动变化(于永翔等, 2023)。但水温变化短期波动较慢,并且在适宜范围内的水温变化对微生物增殖和对虾机体免疫的影响较低且具有滞后性, Pearson 相关分析也表明,当日检测的水温与病原微生物丰度和对虾免疫酶活性的相关性较低。而虾体细菌总数、虾体弧菌占比、水体细菌和盐度 4 个参数的权重高于温度。

目前,国内外许多学者已报道了较多水产动物病害预警模型,大量研究在水质预警方面积累了丰富的经验。其中, Silva 等(2016)构建了基于潮汐、水温和盐度等环境因子的有害藻华预警系统,预测准确性高达 85%。周瑞娟(2011)利用多元线性回归、灰色系统理论和 BP 神经网络等构建了基于水温、透明度和风力等环境因子的大黄鱼(*Larimichthys crocea*)主要细菌性疾病预警模型,其中 BP 神经网络模型准确率最高达到 81.53%。蔡晓鹏(2013)利用多元线性回归、随机森林算法构建了基于水温、溶解氧和化学耗氧量等环境因子的大黄鱼刺激隐核虫病(cryptocaryoniosis)预警预报模型,其中随机森林模型准确率最高达到 83.60%。

水产养殖动物的病原主要以病毒、细菌和寄生虫 3 类为主,目前国内外关于病原相关指标应用于水产病害预警方面的研究报道还较少,其中,冯振飞等(2009)利用多元线性统计方法构建了基于盐度、降雨量和弧菌数的养殖锯缘青蟹(*Scylla serrata*)黄水病预报模型,其准确率达 84.90%,可以较好地用于测报黄水病发病率。宋林生(2020)利用 BP 神经网络构建了基于海水温度、叶绿素含量、弧菌丰度、养殖动物总抗氧化能力和闭壳肌糖原含量的虾夷扇贝病害预警预报体系,其预测效果较好。传统的水产疾病学理论认为,疾病的发生是病原、宿主和环境三者相互影响和作用的结果。笔者认为,病原的存在是疾病发生的重要前提,疾病预警模型体系中包含病原这一指标

是有必要的。本研究构建的基于 Deep Forest 算法的 4 维预警模型的准确率为 89.00%，高于上述没有病原指标的几种预警模型的准确度。

本研究通过对凡纳对虾 AHPND 发生前后的水质因子、健康指标、病原结构三者的协同属性进行分析，建立量化评判指标，解析其变化情况与疾病发生的相关性，并运用多元统计理论和数学演算模型，建立不同参数的相关性方程，确定各个参数在疾病发生过程中的权重大小，进而构建不同疾病的预警数学模型。其中，基于 Deep Forest 算法构建的 AHPND 预警模型准确率较高，并依据因子之间的相关性关系逐步分析筛选，选择以虾体细菌总数、虾体弧菌占比、水体细菌和盐度 4 个参数构建对虾 AHPND 发生预警预测模型。本研究突破了水产养殖系统环境因子、病原数量和宿主健康程度在疾病预警模型中的相关性分析技术，创新水产动物疾病预警的理论体系，为我国水产病害的科学预警和有效防控奠定了技术基础。

参 考 文 献

- BEAZ-HIDALGO R, BALBOA S, ROMALDE J L, *et al.* Diversity and pathogenicity of *Vibrio* species in cultured bivalve molluscs. *Environmental Microbiology Reports*, 2010, 2(1): 34–43
- BI C G, WANG J L, HU N, *et al.* A Bayesian neural network-based early warning model for maize diseases. *Journal of Jilin Agricultural University*, 2021, 43(2): 189–195 [毕春光, 王金龙, 胡楠, 等. 基于贝叶斯神经网络的玉米病害预警模型. 吉林农业大学学报, 2021, 43(2): 189–195]
- CAI X P. Preliminary construction of large yellow croak disease diagnosis and pre-warning models system. Master's Thesis of Xiamen University, 2013 [蔡晓鹏. 大黄鱼疾病诊断与预警系统初步构建. 厦门大学硕士研究生学位论文, 2013]
- CHEN L P, YIN L F, XU T, *et al.* Short-term power load forecasting based on deep forest algorithm. *Electric Power Construction*, 2018, 39(11): 42–50 [陈吕鹏, 殷林飞, 余涛, 等. 基于深度森林算法的电力系统短期负荷预测. 电力建设, 2018, 39(11): 42–50]
- CHEN M M, DONG X, QIU L, *et al.* Quantitative analysis of acute hepatopancreatic necrosis disease causing *Vibrio parahaemolyticus* (V_{AHPND}) in infected *Litopenaeus vannamei*. *Progress in Fishery Sciences*, 2018, 39(4): 93–100 [陈蒙蒙, 董宣, 邱亮, 等. 凡纳滨对虾感染致急性肝胰腺坏死病副溶血弧菌(V_{AHPND})的定量分析. 渔业科学进展, 2018, 39(4): 93–100]
- FENG Z F, WANF G L, NI H E. The forecast model of yellow water disease in mud crab *Scylla serrata*. *Fisheries Science*, 2009, 28(12): 713–716 [冯振飞, 王国良, 倪海儿. 养殖锯缘青蟹黄水病流行病学及其预报模型. 水产科学, 2009, 28(12): 713–716]
- GAO Z G. Research on prediction for quality of transmission and optimization for resource allocation in optical networks based on machine learning. Doctoral Dissertation of Beijing University of Posts and Telecommunications, 2020 [高争光. 基于机器学习的光网络传输质量预测方法与资源优化技术. 北京邮电大学博士研究生学位论文, 2020]
- HA J G, MOON H, KWAK J T, *et al.* Deep convolutional neural network for classifying fusarium wilt of radish from unmanned aerial vehicles. *Journal of Applied Remote Sensing*, 2017, 11(4): 042621
- HAMER W B, BIRR T, VERREET J A, *et al.* Spatio temporal prediction of the epidemic spread of dangerous pathogens using machine learning methods. *International Journal of Geo-Information*, 2020, 9(1): 44
- HAN L, WANG X H, YANG B, *et al.* Analysis of pathogen in an outbreak death of *Marsupenaeus japonicus*. *Journal of Fisheries of China*, 2018, 42(3): 431–441 [韩琳, 王秀华, 杨冰, 等. 一例日本囊对虾暴发性死亡的病原分析. 水产学报, 2018, 42(3): 431–441]
- HUANG L S, ZHANG H S, DING W J, *et al.* Monitoring of wheat scab using the specific spectral index from ASD hyperspectral dataset. *Journal of Spectroscopy*, 2019, 2019(1): 9153195
- JIA D, SHI C Y, HUANG J, *et al.* Identification and pathogenicity analysis of bacterial pathogen associated with acute hepatopancreatic necrosis disease (AHPND) in the Pacific shrimp *Litopenaeus vannamei*. *Progress in Fishery Sciences*, 2018, 39(3): 103–111 [贾丹, 史成银, 黄健, 等. 凡纳滨对虾急性肝胰腺坏死病(AHPND)病原分离鉴定及其致病性分析. 渔业科学进展, 2018, 39(3): 103–111]
- KARVONEN A, RINTAMKI P, JOKELA J, *et al.* Increasing water temperature and disease risks in aquatic systems: Climate change increases the risk of some, but not all, diseases. *International Journal for Parasitology*, 2010, 40(13): 1483–1488
- LIN P C, TANG H. Application of an improved Deep Forest algorithm in insurance purchase prediction scenario. *Modern Information Technology*, 2019, 3(22): 116–122 [林鹏程, 唐辉. 一种改进 Deep Forest 算法在保险购买预测场景中的应用研究. 现代信息科技, 2019, 3(22): 116–122]
- MA Z. The construction of DSS in water quality management of intensive *Litopenaeus vannamei* shrimp tanks. Doctoral Dissertation of Ocean University of China, 2013 [马真. 凡纳滨对虾集约化养殖水质管理决策支持系统的构建. 中国海洋大学博士研究生学位论文, 2013]
- PAILLARD C, LE ROUX F, BORREGO J J. Bacterial disease in marine bivalves, a review of recent studies: Trends and evolution. *Aquatic Living Resources*, 2004, 17(4): 477–498
- RUMPF T, MAHLEIN A K, STEINER U, *et al.* Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Computers and Electronics in Agriculture*, 2010, 74(1): 91–99
- SCHOBER P, BOER C, SCHWARTE L A. Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, 2018, 126(5): 1763–1768

- SCHOFIELD P J, NOBLE B L, CARO L F A, *et al.* Pathogenicity of acute hepatopancreatic necrosis disease (AHPND) on the freshwater prawn, *Macrobrachium rosenbergii*, and Pacific white shrimp, *Penaeus vannamei*, at various salinities. *Aquaculture Research*, 2021, 52(4): 1480–1489
- SILVA A, PINTO L, RODRIGUES S M, *et al.* A HAB warning system for shellfish harvesting in Portugal. *Harmful Algae*, 2016, 53: 33–39
- SONG L S. An early warning system for diseases during mollusc mariculture: Exploration and utilization. *Journal of Dalian Ocean University*, 2020, 35(1): 1–9 [宋林生. 海水养殖贝类病害预警预报技术及其应用. *大连海洋大学学报*, 2020, 35(1): 1–9]
- SOTO-RODRIGUEZ S A, GOMEZ-GIL B, LOZANO-OLVERA R, *et al.* Field and experimental evidence of *Vibrio parahaemolyticus* as the causative agent of acute hepatopancreatic necrosis disease of cultured shrimp (*Litopenaeus vannamei*) in Northwestern Mexico. *Applied Environmental Microbiology*, 2015, 81(5): 1689–1699
- UNDERWOOD K L, RIZZO D M, DEWOOLKAR M M, *et al.* Analysis of reach-scale sediment process domains in glacially-conditioned catchments using self-organizing maps. *Geomorphology*, 2021, 382: 107684
- WAN Q, WANG Q L, WANG R H, *et al.* Short-term load forecasting of a regional power grid based on support vector machine. *Power Grids and Clean Energy*, 2016, 32(12): 14–20 [万强, 王清亮, 王睿豪, 等. 基于支持向量机的某地区电网短期电力负荷预测. *电网与清洁能源*, 2016, 32(12): 14–20]
- XU Y J. Early warning model for water eutrophication based on BP artificial neural network and genetic algorithm. *Computer Science*, 2017, 44(6A): 126–128 [徐云娟. 基于BP神经网络和遗传算法的养殖水域预警模型. *计算机科学*, 2017, 44(6A): 126–128]
- YU Y X, WANG Y G, CAI X X, *et al.* The occurrence of AHPND in pond cultured *Litopenaeus vannamei* and its influence on environment, pathogen and immune factors. *Acta Hydrobiologica Sinica*, 2023, 47(1): 1–10 [于永翔, 王印庚, 蔡欣欣, 等. 环境、病原、免疫因子三要素与池塘养殖对虾 AHPND 发生的关联性. *水生生物学报*, 2023, 47(1): 1–10]
- ZHANG J L. Research on water quality factors and aquatic organism in *Litopenaeus vannamei*. Master's Thesis of Shanghai Ocean University, 2016 [张健龙. 凡纳滨对虾养殖池水质因子及水生生物研究. 上海海洋大学硕士研究生学位论文, 2016]
- ZHOU R J. Studies on the forecasting of the main disease in cage-cultured *Pseudosciaena crocea*. Master's Thesis of Ningbo University, 2011 [周瑞娟. 网箱养殖大黄鱼主要疾病的预报模型研究. 宁波大学硕士研究生学位论文, 2011]

(编辑 马瑾艳)

Construction of an Early Warning Mathematical Model for *Penaeus vannamei* AHPND Based on the Deep Forest Algorithm

WANG Yingeng^{1,2}①, YU Yongxiang^{1,2}, CAI Xinxin¹, ZHANG Zheng^{1,2},
WANG Chunyuan¹, LIAO Meijie^{1,2}, ZHU Hongyang¹, LI Hao¹

(1. Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao 266071, China;

2. Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao Marine Science and Technology Center, Qingdao 266071, China)

Abstract Acute hepatopancreatic necrosis disease (AHPND) is widely prevalent, has a rapid onset, and has high mortality in shrimp culture, making it a key limiting factor affecting shrimp aquaculture development in recent years, resulting in massive economic losses to the industry worldwide. Systematic studies that investigate which factors significantly correlate with the occurrence of AHPND, and further establishment of a prediction model for the occurrence of shrimp AHPND, are important for preventing and controlling the disease. In this study, *Penaeus vannamei* in pond culture were preliminarily analyzed for the coupling relationship between the occurrence and prevalence of AHPND in shrimps and pathogens, and for environmental and host autoimmune factors by assessing the environmental factors, pathogen

① Corresponding author: WANG Yingeng, E-mail: wangyg@ysfri.ac.cn

abundance, and host health indicators under AHPND incidence. Then, a mathematical early warning model of AHPND occurrence in pond-cultured *P. vannamei* was constructed using Deep Forest algorithm. The occurrence of AHPND and its environment, pathogen, and shrimp immunity factors in pond-cultured *P. vannamei* were continuously monitored to explore the relationship between the occurrences of AHPND in relation to these factors. A total of 18 parameters were assessed, including the air and water temperature, salinity, pH, dissolved oxygen (DO), ammonia nitrogen ($\text{NH}_4^+\text{-N}$) and nitrite ($\text{NO}_2\text{-N}$) concentrations, culturable bacteria and *Vibrio* in water, culturable bacteria and *Vibrio* in the shrimp hepatopancreas, the proportion of *Vibrio* in water and the shrimp hepatopancreas, and the activities of acid phosphatase (ACP), alkaline phosphatase (AKP), superoxide dismutase (SOD), lysozyme (LZM), and phenol oxidase (PO) in shrimp muscles. The parameter simulation prediction data based on the *P. vannamei* AHPND occurrence-related factor sequence (environmental factor, microbial factor, and shrimp health indicator) were constructed for the first time. The one-dimensional sequence was mapped into the three-dimensional space, different kernel functions were selected in combination with the actual classification problem to compare the model fitting accuracy, and the test algorithm optimized the parameters in the model. A total of 140 relevant data groups were collected under the same mode, and the groups of additional exogenous inputs during the breeding process were eliminated. After deleting invalid data, there were 100 groups of classified monitoring data, including 25 groups of morbidity data and 75 groups of health data. Moreover, the model was affected due to the dimensional and quantitative differences among different factors. In order to improve the speed of subsequent experimental training and prediction accuracy, the 100 groups of training test data processed by the mapminmax function were normalized for data processing. The relationship between 18 parameters and the occurrence of AHPND in *P. vannamei* was analyzed using Pearson's correlation, and the main influencing factors were further screened using pairwise analysis between the factors. Pearson's correlation analysis indicated that the incidence of AHPND positively correlated ($P < 0.05$) with salinity, the number of culturable bacteria and *Vibrio* in the shrimp, the proportion of *Vibrio* in the shrimp, the number of culturable bacteria and *Vibrio* in water, and the activities of LZM, ACP, and PO in shrimp muscles. The correlation coefficients were 0.350 1, 0.574 1, 0.521 1, 0.391 1, 0.374 7, 0.238 3, 0.438 2, 0.257 1, and 0.228 9, respectively, indicating that AHPND was more likely to occur with an increase of these parameter values within a certain range. The incidence of AHPND negatively correlated with water temperature ($P < 0.05$), and the correlation coefficient was $-0.227 9$. Moreover, the water temperature, pH, DO, $\text{NH}_4^+\text{-N}$ and $\text{NO}_2\text{-N}$ concentrations, *Vibrio* proportion in water, AKP, and SOD had a weak correlation with the incidence of AHPND ($P > 0.05$). Furthermore, parameters were removed in the model construction process according to the correlation between parameters and factors. The occurrence of AHPND in *P. vannamei* directly and significantly correlated with seven parameters, including the total number of shrimp bacteria, the total number of shrimp *Vibrio*, LZM, the proportion of shrimp *Vibrio*, the total number of water bacteria, salinity, and the total number of water *Vibrio*. The prediction performance of three popular integrated learning method algorithms based on decision tree, Deep Forest, LightGBM, and XGBoost was evaluated using Python language programming, and, finally, a four-dimensional vector early warning prediction model based on the Deep Forest algorithm for the total number of shrimp bacteria, the proportion of *Vibrio* shrimp, the total number of water bacteria, and salinity was established (accuracy: 89.00%). Although the prediction performance of the Deep Forest model decreased somewhat compared with that of the support vector machine model established in this study, the algorithm was gradually screened out based on the correlation between factors, including the effects of all factors. It was proven that the Deep Forest model established in this study was the ideal prediction model for predicting the occurrence of AHPND in *P. vannamei* among the 10 dimension parameters tried, and the superiority of the Deep Forest algorithm was also further verified. The results provide basic data and technical support for shrimp AHPND disease prediction, prevention and control, and lay a theoretical foundation for further establishment of aquaculture animal disease early warning theory.

Key words Shrimp; AHPND; Early warning mathematical model; Deep Forest algorithm; Python programming language