·调查报告与分析·

输入性登革热预警模型构建：基于多融合机器学习法

赵婕[1]，顾大勇[2]，张建庆[3]，吴育凡[4]，文洁[5]，江婷婷[1]，王宇平[1,3]

1. 福建医科大学公共卫生学院，福州 350004；
2. 深圳大学，深圳 518000；
3. 福州国际旅行卫生保健中心，福州 350001；
4. 深圳市第二人民医院，深圳 518000；
5. 广西中医药大学，南宁 530200
通信作者：王宇平，E-mail：ericwangyp@hotmail.com

【摘　要】目的 本研究旨在构建基于多融合机器学习的输入性登革热预测模型，以提升广州口岸传染病防控的科学性和前瞻性。**方法** 收集 2015—2019 年广州口岸输入性登革热病例数，并结合气象、入境旅客、航班流量、植被指数和媒介密度等多源数据，进行清洗与标准化处理后，分别建立 BP 神经网络、随机森林和 CatBoost 模型，并在此基础上采用加权平均、堆叠集成等方式构建双融合和三融合模型。模型性能通过均方误差（MSE）和决定系数（R^2）评估。**结果** 融合模型整体优于单一模型，表现出更高的预测精度与稳定性。其中，随机森林与 CatBoost 的堆叠融合模型表现最佳，MSE 为 4.749，R^2 提升至 0.601，较最佳单一模型显著改善。SHAP 分析进一步揭示了湿度、日照时数及昼夜温差等气象因子为主要驱动变量，入境旅客数量亦对预测结果有重要贡献。**结论** 多融合机器学习模型在输入性登革热预测中具有显著优势，可为口岸公共卫生早期预警与防控措施的制定提供有力支撑。

【关键词】 登革热；机器学习；预测模型；公共卫生；广州市

Construction of an early warning model for imported dengue fever: a multi-fusion machine learning approach

ZHAO Jie[1], GU Dayong[2], ZHANG Jianqing[3], WU Yufan[4], WEN Jie[5], JIANG Tingting[1], WANG Yuping[1,3] (*1. School of Public Health*, *Fujian Medical University*, *Fuzhou* 350004, *China*; *2. Shenzhen University*, *Shenzhen* 518000, *China*; *3. Fuzhou International Travel Healthcare Center*, *Fuzhou* 350001, *China*; *4. Shenzhen Second People's Hospital*, *Shenzhen* 518000, *China*; *5. Guangxi University of Chinese Medicine*, *Nanning* 530200, *China*)
Corresponding author: WANG Yuping, E-mail: ericwangyp@hotmail.com

【Abstract】 Objective To develop an imported dengue fever prediction model based on multi-fusion machine learning, aiming to enhance the scientific basis and foresight of infectious disease prevention and control at Guangzhou port. **Methods** Data on imported dengue fever cases at Guangzhou port from 2015 to 2019 were collected, along with meteorological indicators, inbound passenger volume, flight traffic, vegetation index, and vector density. After data cleaning and standardization, three base models (BP neural network, random forest, and CatBoost) were constructed. Weighted averaging and stacking ensemble strategies were then adopted to build dual-fusion and triple-fusion models. Model performance was evaluated based on mean squared error (MSE) and coefficient of determination (R^2). **Results** The fusion models outperformed single models overall, demonstrating higher predictive accuracy and stability. Among them, the stacked ensemble of random forest and CatBoost achieved the best performance, with an MSE of 4.749 and an R^2 of 0.601, significantly improving upon the best single model. SHAP analysis further revealed that humidity, sunshine duration, and diurnal temperature variation were the main driving factors, while inbound passenger volume also contributed substantially to predictions. **Conclusions** Multi-fusion machine learning models provide significant advantages in predicting imported dengue cases, offering strong support for early warning and the formulation of prevention strategies in port public health.

【Keywords】 dengue fever; machine learning; prediction model; public health; Guangzhou

开放获取：CC BY-NC-ND 4.0 DOI：10.11847/zgggws1148778
基金项目：国家重点研发计划（2022YFC2302700）
第一作者：赵婕（2000—），硕士，研究方向：疾病预防与控制。
收稿日期：2025-09-01 修回日期：2025-11-26 录用日期：2026-03-07
利益冲突：不存在 伦理审查：免除审查 出版授权：全体作者已与编辑部签署作者贡献声明及版权转让协议

登革热（dengue fever, DF）是由登革病毒引起、主要通过埃及伊蚊（*Aedes aegypti*）和白纹伊蚊（*Aedes albopictus*）传播的虫媒病毒性疾病，广泛分布于美洲、东南亚、西太平洋等全球热带和欧亚热带地区，近年对全球健康造成威胁[1]。登革热自 1978 年在广东省佛山市暴发以来，在我国间断性流行，流行的趋势为每隔 4～7 年发生一次[2]，广东省又以广州和佛山市最为突出[3]。广州市地处广东省的中南部、珠江三角洲的北缘，接近珠江流域下游入海口，其范围为东经 112°57′～114°3′、北纬 22°26′～23°56′[4]。广州市是中国远洋航运的优良海港和珠江流域的进出口岸，又是京广、广深、广茂和广梅汕铁路的交汇点和华南民用航空交通中心，有中国“南大门”之称[5]。广州市海洋性气候特征显著，具有温暖多雨、光热充足、温差较小、夏季长、霜期短等特征，为蚊媒繁殖及登革热传播提供了适宜的生态环境[6]。加强登革热疫情的预测预警研究，利于积极应对日益严峻的登革热疫情。近年来，随着数据多样性和计算能力的提升，机器学习方法因其强大的特征处理能力和灵活建模结构，在传染病预测领域得到广泛应用，例如深度学习模型用于气候与空间变量联合预测登革热暴发[7]。多模型融合（multi-model ensemble）成为提升预测性能的一种有效策略。例如，在巴西的全国级别登革热预测中，Sebastianelli A 等[8] 通过将 CatBoost、支持向量机（SVM）和长短时记忆网络（LSTM）结合为一个融合模型，显著提升了预测准确性。此外，在越南湄公河三角洲地区的登革热预测研究中，融合气象、社会人口和半机制模型的集成架构相比单一模型提升了近 69% 的预测精度[9]。本研究选取广州口岸输入性登革热病例为预测对象，整合气象、旅客流量与媒介监测等多源数据，构建 BP 神经网络、随机森林与 CatBoost 这 3 类模型，并探索加权平均与堆叠融合等多种融合策略，旨在评估这些融合模型在输入性登革热预测中的性能表现与解释力，为口岸传染病早期预警提供科学支持。

1 资料与方法

1.1 数据采集

1.1.1 流行病学数据　研究数据涵盖 2015—2019 年广州口岸的输入性病例信息，病例来源于海关总署系统，包括广州白云国际机场海关、南沙海关、广州车站海关及中山局本部。同时选取登革热输入病例数最多的 3 个国家（印度尼西亚、马来西亚和越南）作为重点输入国，采集同期登革热病例数据。数据来源包括世界卫生组织疫情周报（WHO Weekly Epidemiological Record）、全球疫情警报与反应网络（GOARN）及目标国家卫生部门公告等。通过跨语言信息抽取和地理空间映射，建立结构化的输入国疫情数据库。

1.1.2 气候数据　温度、湿度和降雨量等气象因素主要通过改变蚊虫的分布范围，增加蚊虫的孳生场所、繁殖速度与侵袭力等方式影响登革热的传播。本研究重新整理 2015—2019 年广州地区月平均气温、月平均相对湿度等气象因素并建库；国内数据来源于中国科学院资源环境科学数据中心（http://www.resdc.cn）的气象数据，全球气象数据来源于美国国家海洋和大气管理局（NOAA）下设的国家环境信息中心（NCEI）发布的气象数据集。

1.1.3 虫媒密度数据　本研究数据于 2023 年 3 月获取，数据来源于广州口岸开展的虫媒密度常规监测。监测范围涵盖白云机场海关、南沙海关和广州车站海关，数据时间范围为 2015 年至 2019 年。根据不同生境类型，将虫媒密度指标划分为室内密度和室外密度，并按照《全国病媒生物监测方案》的相关要求，实施各口岸环境月度监测。采用二氧化碳灯诱法和诱卵器法监测蚊类，蚊媒密度计算公式为：平均密度［只/（灯·小时）］=捕蚊总数/（诱蚊灯数×2）；诱蚊诱卵指数（%）=［（阳性诱卵器数/布放诱卵器总数）］×100%。

1.1.4 社会经济数据和交通流量数据　包括航班数、人流量、人口密度等来源于国家统计局（http://www.stats.gov.cn）。广东每日的国际航班数据来源于 VariFlight 网站（http://www.variflight.com），数据包括航班日期、航班号、出发国家、到达机场、到达城市、实际到达时间、最大座位数；广东省人口数据和地区面积数据来源于《广东统计年鉴 2020》，东南亚国家人口数据来源于世界卫生组织（https://ntdhq.shinyapps.io/dengue5/）。

1.2 方法

1.2.1 统计学处理　原始数据经 Excel 2024 软件汇总和处理并建库。为了消除量纲差异对模型性能的影响、增强数据之间的可比性，对数据集进行了标准化处理和缺失值处理，保证特征分布的合理性并减少极端观测对模型拟合的干扰。为建立稳定的预测模型，划分为训练集和测试集，并通过机器学习方法提取训练集特征值和目标值进行预测，基于测试集确定模型训练精度，将数据集以 8:2 的比例划分为训练集和测试集，确保模型能够在未见数据上的正确学习，以有效评估模型稳定性。

1.2.2 模型建立和融合方法　在模型选择方面，研究中使用的 3 类基础学习器包括 BP 神经网络、随机森林和 CatBoost 梯度提升树。三者分别侧重于不同的拟合能力，BP 神经网络能够表征复杂的非线性关系，随机森林具有较强的特征交

互捕捉能力，而 CatBoost 在处理异质性变量时表现稳健。为了进一步提升预测精度，本研究在单模型的基础上设计了多种融合策略。双模型层面上，分别对 BP 与随机森林、BP 与 CatBoost、随机森林与 CatBoost 进行加权平均融合；三模型层面上，则尝试了加权平均与堆叠式集成等方法，并通过时间序列交叉验证确定融合权重，以最小化均方误差作为优化目标。本研究使用的融合方法有加权平均法和堆叠法（stacking），可以用来提高多个不同模型组合的预测准确性。它通过将多个基本模型（也称为第一层模型）的预测结果作为输入，来训练一个高级模型（也称为第二层模型或元模型），以此实现更优的综合预测效果。堆叠法的核心策略是利用多个不同的基模型对数据集的不同方面或不同方式的理解能力，然后通过一个元模型来综合这些基模型的预测结果。堆叠法的元模型能够学习如何最优地结合各个基模型的预测，以利用各个模型的强项，弥补各个模型的不足。

1.2.3 模型评价 模型建立后，采用校验样本对模型效果进行验证。本文所研究的传染病预测模型主要通过均方误差（mean squared error, MSE）和决定系数（coefficient of determination, R^2）评估预测性能。

2 结 果

2.1 模型评估结果 在上述数据集及模型理论的基础上，本研究在 PyCharm 平台上采用 Python 3.9 建立输入性登革热预测模型，分别建立了 RF、BP、CatBoost 模型以及基于 Stacking 方法的多模型决策级融合模型。为进一步表现决策级模型的准确率，将单一模型计算得到的结果与融合模型进行对比分析。在本研究中，多个模型对广州口岸输入性登革热病例的预测性能表现出显著差异。总体来看，融合模型在多数情境下优于单一模型，验证了多源集成方法在复杂流行病学数据建模中的有效性。其中，随机森林与 CatBoost 的堆叠融合模型（DualFuse-StackLR-RFCB）表现最优，其均方误差最低（4.749），决定系数达到 0.601，表明该模型不仅在误差控制上具备优势，而且对病例波动趋势的解释力亦最为充分。BP 神经网络单模型的预测精度次之，均方误差为 5.153，决定系数为 0.568，显示其在捕捉非线性关系方面依然具备较强潜力（表 1）。从融合效果来看，双模型平均融合与加权融合的结果虽较单一的 CatBoost 模型有所改善，整体仍低于 BP 神经网络和堆叠融合模型，简单加权在本研究中提升有限，而更为复杂的堆叠策略能够更有效地整合不同模型的优势。另外，堆叠融合在模型性能上的优势源于其能够通过二层学习器自适应地分配各基学习器的贡献权重，从而显著提高整体预测的稳定性和精度。

表 1 不同模型性能评估参数
Table 1 Model performance evaluation parameters

模型	MSE	R^2值
DualFuse-StackLR-RFCB	4.749	0.601
BP-NeuroNet	5.153	0.568
DualFuse-AVG-BPCB	5.593	0.531
DualFuse-WGT-BPCB	5.671	0.524
CatBoost	6.554	0.450
TriFuse-WGT-BCR	6.930	0.418
TriFuse-AVG-BCR	6.937	0.417
DualFuse-AVG-BPRF	7.298	0.388
DualFuse-WGT-BPRF	7.349	0.383
TriStack-LR-BCR	7.698	0.354
DualFuse-WGT-RFCB	8.025	0.327
DualFuse-AVG-RFCB	8.180	0.314
DualFuse-BoostStack-RFCB	8.436	0.292
RandomForest	10.145	0.149
TriStack-GB-BCR	15.243	–0.278
DualFuse-BoostStack-BPCB	15.398	–0.292
DualFuse-StackLR-BPCB	15.428	–0.295
DualFuse-StackLR-BPRF	17.121	–0.437
DualFuse-BoostStack-BPRF	18.584	–0.559

2.2 模型拟合效果可视化 图 1 展示了不同模型预测值与真实值之间的拟合关系。以 BP-NeuroNet 为代表的单一模型能够较好地捕捉输入性登革热病例的变化趋势，其预测点大体分布在理想拟合线（45°线）附近，显示出较强的线性一致性。相比之下，融合模型在拟合精度上表现更为优越。例如，DualFuse-AVG-BPCB 有效减弱了单模型预测中的偏差，使得预测点更加紧密地聚合在理想拟合线周围，说明融合方法能够在不同基学习器之间实现优势互补，从而提升模型稳健性和准确性。DualFuse-StackLR-RFCB 通过堆叠策略整合随机森林与 CatBoost 的预测结果，在整体拟合趋势上表现更佳，验证了多模型集成在输入性登革热预测中的潜在优势。

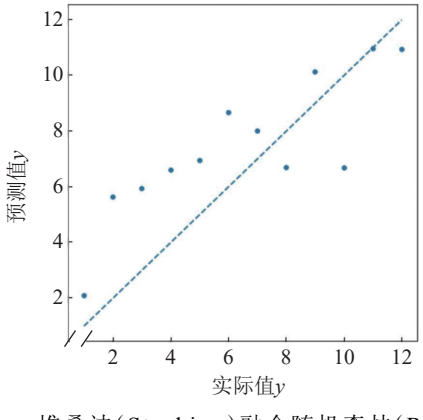


图 1a 堆叠法（Stacking）融合随机森林（RF）和 CatBoost（CB）散点图
Fig. 1a Prediction performance of DualFuse-StackLR-RFCB

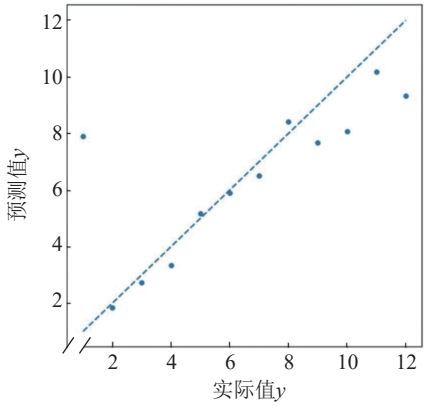


图 1b　BP 神经网络模型散点图
Fig. 1b　Prediction performance of BP-NeuroNet

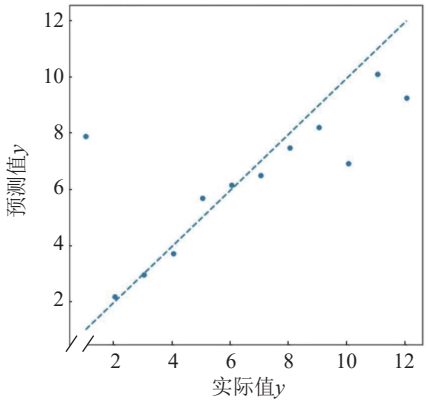


图 1c　平均法(AVG)融合神经网络(BP)和CatBoost(CB)散点图
Fig. 1c　Prediction performance of DualFuse-AVG-BPCB

2.3　特征贡献度和解释

2.3.1　基于最佳模型的 SHAP 值汇总分析(图 2)

图 2 横坐标为 SHAP 值大小，反映了各变量对模型预测的影响强度，纵坐标则为输入特征，颜色代表特征取值的高低。整体来看，相对湿度是影响模型预测的最主要因素，SHAP 值分布范围最广，表明湿度水平的波动对输入性登革热风险具有显著驱动作用。湿度的 SHAP 贡献最为分散，提示湿度对病毒传播可能存在复杂的非线性关系。在气象因素中，日照时数与昼夜温差比值紧随其后，二者 SHAP 值均显示了较为明显的解释力，说明光照条件与温度波动对蚊媒繁殖及病毒传播风险均具有关键作用。平均气温、最高气温与最低气温亦呈现出重要贡献，强化了气温作为登革热输入风险核心驱动因子的作用。除气象因素外，入境旅客数量和航班数量也在模型中展现出一定影响，其中入境旅客数量在非气象因素中贡献最高，提示人口流动仍然是跨境输入病例的潜在因素。与此同时，来自越南、马来西亚与印尼的输入病例数在模型中呈现中等水平的重要性，且不同来源地的特征取值方向性有所差异，表明病例来源国的差异性在模型预测中不可忽视。其他变量如植被指数(NDVI)、累计降雨量、蚊密度在整体预测中的解释力相对较低，但在局部样本中仍可能起到修正作用。该 SHAP 汇总分析表明，气象条件(尤其是湿度与气温)、人口流动规模以及部分输入病例来源国是驱动输入性登革热预测的核心变量。

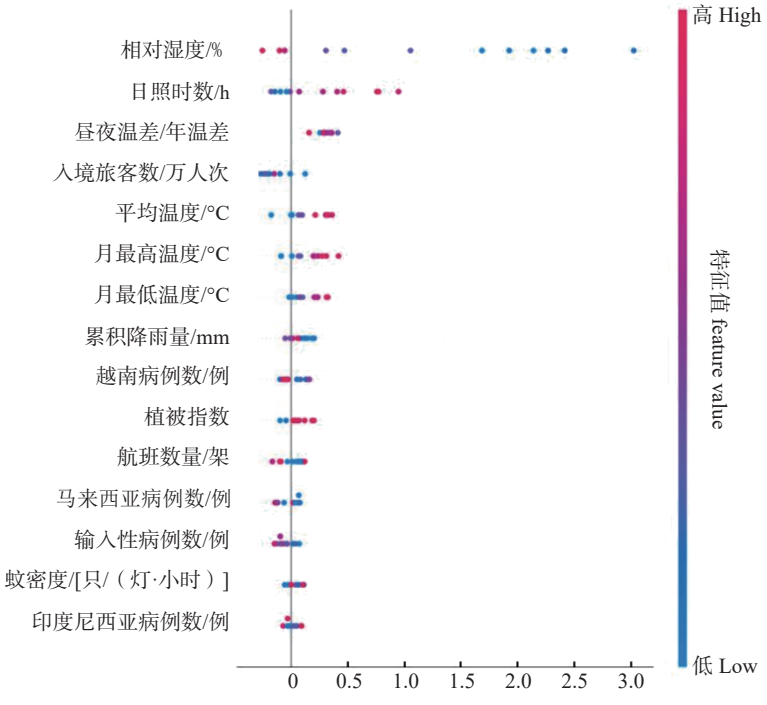


图 2　SHAP 图
Fig. 2　SHAP summary plot

2.3.2 基于最佳模型的特征重要性条形图(图 3)结果显示，相对湿度对模型预测的贡献最大，其平均 SHAP 值远高于其他变量，表明湿度是决定输入性登革热风险预测的首要驱动因子。紧随其后的是日照时数与昼夜温差比值，说明气候条件在模型中占据核心地位，并对登革热风险预测产生系统性影响。平均气温、最高气温与最低气温同样位居前列，进一步强化了温度相关因素在输入性登革热传播风险预测中的决定性作用。除气象变量外，入境旅客数量在所有非气象因素中贡献最高，表明跨境人口流动对输入病例风险预测也具有显著影响。相比之下，越南、马来西亚与印尼病例输入数的整体贡献度相对有限，但仍在模型中发挥补充作用，为病例来源地差异提供了解释。累计降雨量、航班数量、NDVI 指数与蚊密度在模型中的贡献相对较低，提示上述指标可能更多地作为次要修正因素存在。

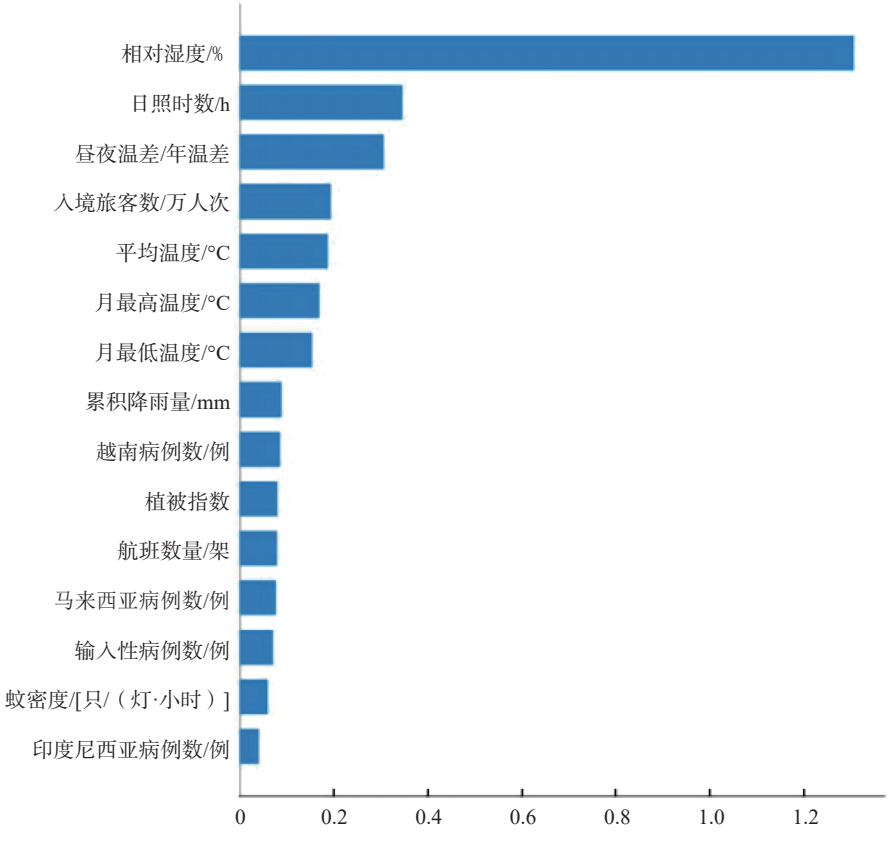


图 3 基于最佳模型的特征重要性排序

Fig. 3 Feature importance ranking based on the optimal model

3 讨 论

本研究构建的多融合机器学习模型在预测广州口岸输入性登革热方面展现出显著优势，相较于简单加权融合，本研究在输入性登革热这一受多源异质因素共同驱动的预测场景中引入二级学习器的堆叠结构，更有利于整合模型的互补信息，提升预测结果的稳定性与实际应用价值，这与全球范围内对登革热传播预测模型的研究趋势保持一致。例如，Sebastianelli A 等人[8] 以巴西为研究对象，构建了基于 CatBoost、SVM 和 LSTM 的机器学习融合预测框架，结果表明，该集成模型在全国尺度的登革热发病率预测中整体优于单一模型，并在跨国家数据集上的验证中显示出一定的可迁移性。在多模型对比研究中，Chen X[10] 以巴西里约热内卢为研究对象，对多种统计模型与机器学习模型的登革热预测性能进行了系统评估。研究结果表明，机器学习模型在多数预测情景下优于单一模型，尤其在引入滞后气象变量的预测窗口中表现出更高的稳定性与预测精度[10]。该结论与本研究中融合模型相较单一模型在 MSE 和 R^2 指标上的显著改进一致，表明融合策略能够有效提升传染病预警模型的预测性能。

本研究运用 SHAP 分析量化了气象因子与跨境人口流动等变量的重要性。研究发现，相对湿度、日照时数和温差等气象因素是最重要的驱动因子，与 Zhao N Z 等[11] 对矢量指数与气象数据在登革热传播率预测中的结果一致，说明气候条件在预测中的核心作用；此外，Zhao N Z 等还在哥伦比亚的全国与地方模型中发现在短期预测中环境变量尤为关键，而社会人口统计因素则对

长期预测更具影响力。此证据支持了本研究中相对湿度较高时具有负向贡献，而人口流动变量在一定程度上增强输入预测能力的结果。在方法创新上，本研究采用的 3 模型融合方法（包括加权融合与堆叠层融合）体现了融合学习的优势。Panja M 等[12] 提出了一种融合小波分解与神经网络的集成预测模型（XEWNet），通过刻画气候变量的多尺度时序特征，在多个地区的登革热预测中较单一模型和传统方法表现出更高的预测精度与稳定性。Draidi Areed W 等人[9] 在越南湄公河三角洲地区构建了融合时空信息与统计模型的区县级登革热预测系统，该模型在 1～3 个月预测窗口内的预测准确率约为 69%。该研究表明，基于多模型融合的预测方法在中短期登革热风险评估中具有较好的应用潜力，为区域防控的决策提供了支持。

本研究仍存在一些局限性及未来改进方向。首先，数据变量尚可进一步扩展至媒介指数、社交媒体数据等，Guo P 等人[13] 曾在广州使用百度搜索和微博数据结合气象与病例信息，构建了融合性强的预警模型。此外，本研究未充分考虑免疫力或历史流行态势，这在某些研究中被证明能显著提升预测准确性[14]。未来研究可通过引入宿主免疫监测指标、登革病毒不同血清型或基因型信息，以及媒介密度等变量来丰富模型输入特征，以进一步提升模型的解释性与预测稳定性。此外，模型在不同国家或流行病学背景下的泛化能力仍有待通过多中心、多环境数据加以验证。如 Al Mobin M[15] 在孟加拉国采用时序数据下采样与 ML 建模框架，优化了预测准确率，提供了一种迁移可行性的研究方法。

数据可用性声明　研究分析的数据来源于中国海关总署系统，受版权协议限制，仅限授权使用，可联系该机构申请访问权限（网站地址：http://www.customs.gov.cn/）。本研究所有观点和结论均属作者观点，不代表中国海关总署意见或政策。

参考文献

[1] Kularatne S A, Dalugama C. Dengue infection: global importance, immunopathology and management[J]. Clinical Medicine, 2022, 22(1): 9-13.

[2] Yue Y J, Liu X B, Ren D S, et al. Spatial dynamics of dengue fever in China′s mainland, 2019[J]. International Journal of Environmental Research and Public Health, 2021, 18(6): 2855.

[3] 邹钦. 广东省 2014 年登革热疫情防控实践与思考 [J]. 中华卫生杀虫药械, 2015, 21(6): 630-632.

[4] 广州市人民政府. 自然地理 [EB/OL]. (2024-12-29) [2025-07-08]. https://www.gz.gov.cn/zlgz/gzgk/zrdl/content/post_10071227.html.

[5] 邹海滨, 宋卫, 林厚炳, 等. 珠江水域广州口岸蚊类本底调查 [J]. 中国国境卫生检疫杂志, 2014, 37(5): 331-335.

[6] 周金华, 何时雨, 刘通, 等. 2021—2023 年广州市白纹伊蚊监测 [J]. 中国血吸虫病防治杂志, 2025, 37(1): 76-80.

[7] Chen X, Moraga P. Forecasting Dengue across Brazil with LSTM neural networks and SHAP-driven lagged climate and spatial effects[J]. BMC Public Health, 2025, 25(1): 973.

[8] Sebastianelli A, Spiller D, Carmo R, et al. A reproducible ensemble machine learning approach to forecast dengue outbreaks[J]. Scientific Reports, 2024, 14(1): 3807.

[9] Draidi Areed W, Nguyen T T T, Quoc Do K, et al. A district-level ensemble model to enhance Dengue prediction and control for the Mekong Delta Region of Vietnam[J]. PLoS Neglected Tropical Diseases, 2025, 19(9): e0013571.

[10] Chen X, Moraga P. Assessing dengue forecasting methods: a comparative study of statistical models and machine learning techniques in Rio de Janeiro, Brazil[J]. Tropical Medicine and Health, 2025, 53(1): 52.

[11] Zhao N Z, Charland K, Carabali M, et al. Machine learning and Dengue forecasting: comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia[J]. PLoS Neglected Tropical Diseases, 2020, 14(9): e0008056.

[12] Panja M, Chakraborty T, Nadim S S, et al. An ensemble neural network approach to forecast Dengue outbreak based on climatic condition[J]. Chaos, Solitons and Fractals, 2023, 167: 113124.

[13] Guo P, Zhang Q, Chen Y L, et al. An ensemble forecast model of Dengue in Guangzhou, China using climate and social media surveillance data[J]. Science of the Total Environment, 2019, 647: 752-762.

[14] McGough S F, Clemente L, Kutz J N, et al. A dynamic, ensemble learning approach to forecast Dengue fever epidemic years in Brazil using weather and population susceptibility cycles[J]. Journal of the Royal Society Interface, 2021, 18(179): 20201006.

[15] Al Mobin M. Forecasting dengue in Bangladesh using meteorological variables with a novel feature selection approach[J]. Scientific Reports, 2024, 14(1): 32073.