医疗大数据在公共卫生监测与预警中的应用

张志超

天津市斯宇克医疗器械销售有限公司　天津　300000

摘　要：目的：本文旨在构建结合医疗大数据的多层次公共卫生监测与预警体系，以提高疾病传播趋势预测、提前预警及异常检测的效率和精度，为公共卫生管理提供科学依据。方法：基于 SEIR 模型、LSTM 模型和 Isolation Forest 模型，分别实现疾病传播动态模拟、趋势预测及异常检测。结果：LSTM 模型在预测精度上表现最佳，其 RMSE 为 10.67，MAPE 为 5.1%，提前预测天数达到 7 天。SEIR 模型成功模拟了流感传播趋势，参数估算值 $\beta=0.12, \sigma=0.20, \gamma=0.15$，与实际病例数据对比的 RMSE 为 14.25。Isolation Forest 模型的检测准确率达到 96.2%，召回率为 93.5%，有效识别了流感暴发的异常信号。结论：结合医疗大数据的多层次建模体系显著提升了公共卫生监测与预警的效率与精度。通过进一步优化数据该体系可为流行病的早期干预和政策制定提供更科学的支持。

关键词：医疗大数据；公共卫生监测；疾病传播预测；提前预警

1. 引言

近年来，公共卫生事件的频发以及全球化进程的加速，使得公共卫生监测与预警成为全球卫生治理的重要议题。传统的监测与预警手段依赖于有限的监测点和定期的数据汇报，难以及时、全面地捕捉公共卫生风险的动态变化。而医疗大数据的迅速发展为构建高效、精准的公共卫生监测与预警体系提供了新的技术支持。医疗大数据来源广泛，涵盖电子病历、流行病学调查、基因检测和健康管理平台等，其高维度性、时空相关性与动态性为传染病传播建模和异常趋势检测提供了丰富的信息资源。通过合理利用医疗大数据，可以有效提升疾病监测的广度与深度，同时基于时空动态实现早期预警，为公共卫生政策的制定提供科学依据。本文以医疗大数据为核心，构建适应实际需求的多层次监测与预警模型，并通过实证分析验证其效果与应用价值。

2. 相关理论分析

2.1 医疗大数据的特性与作用

医疗大数据是现代信息技术与医学结合的产物，其核心特性决定了其在公共卫生监测与预警中的应用优势 [1]。第一，医疗大数据具有高维度性，涵盖患者的临床特征、环境变量（如气温、湿度）、行为模式（如人群流动性）等多维特征，使得数据分析能够捕捉复杂的健康与环境交互关系，为疾病传播预测提供多元信息支持 [2]。第二，其动态性体现在实时更新的特性，这有助于捕捉公共卫生事件的动态变化，通过时序数据分析发现疾病传播的早期信号，从而有效降低大规模暴发的风险 [3]。第三，医疗大数据还具备时空相关性，能够结合地理信息系统（GIS）技术，实现疾病传播的时空动态监测，为精准的区域性防控提供科学依据 [4]。最后医疗大数据的大规模与多源性使其整合了医院、社区卫生服务、传染病监测系统等异构数据，支持宏观监测与细粒度分析相结合，提升了公共卫生监测的全局性与细致性。

2.2 公共卫生监测与预警理论

公共卫生监测与预警的理论基础涵盖传染病传播建模、时空动态监测及异常检测三大关键方面 [5]。SEIR 模型通过描述易感者、潜伏者、感染者和康复者的动态变化，能够模拟疾病传播的过程。结合医疗大数据，可以动态调整关键参数，提高模型对实际传播规律的适应性，从而实现精准的疫情趋势预测。不同地区的疾病传播速度和风险具有显著差异，通过地理加权回归模型与 GIS 技术相结合的时空动态监测方法，能够捕捉疫情扩散的时空异质性，支持针对性区域防控决策。为实现疾病暴发的早期预警，异常检测技术至关重要。以 Isolation Forest 为代表的异常检测方法，通过评估病例数据的异常点，能够快速识别潜在的公共卫生风险信号。以上三类模型从不同维度协同构建了基于医疗大数据的多层次公共卫生监测与预警体系，有效支持传染病的早期发现和精准防控。

3. 模型

3.1 传染病传播模型

传染病传播建模是公共卫生监测与预警的核心之一，用于描述疾病传播的动态过程。SEIR 模型（易感者－潜伏者－感染者－康复者模型）是一种经典的传染病传播模型，其数学描述如下：

$$\frac{dS(t)}{dt} = -\beta S(t)I(t),$$
$$\frac{dE(t)}{dt} = \beta S(t)I(t) - \sigma E(t),$$
$$\frac{dI(t)}{dt} = \sigma E(t) - \gamma I(t),$$
$$\frac{dR(t)}{dt} = \gamma I(t),$$

其中 S(t) 是时刻易感者比例；E(t) 是潜伏者比例；I(t) 是感染者比例；R(t) 是康复者比例；β 是疾病传播速率；σ 是潜伏期向感染期的转化速率；γ 是感染者康复速率。通过引入医疗大数据，可以动态调整 β，σ，γ 的值，以反映环境变量（如气温、湿度）和人群流动性的影响。相关公式如下。

$$\beta = \beta_0 + \beta_1 T + \beta_2 H + \beta_3 M,$$

其中 T 为气温，H 为湿度，M 为人群流动性指数。该模型能够根据实时数据预测疾病传播趋势，帮助制定精准的公共卫生干预措施。

3.2 大数据驱动的预测模型

基于医疗大数据的预测模型在公共卫生监测中扮演了重要角色，用于捕捉复杂数据模式并预测未来的疾病发展。长短期记忆网络 (LSTM) 是一种常用的深度学习模型，适用于时间序列数据分析。LSTM 的核心数学结构如下。

1. 遗忘门

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f),$$

f_t 控制历史信息的遗忘比例，W_f 是权重矩阵，b_f 是偏置项

2. 输入门

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i),$$

i_t 决定当前输入的记忆重要性。

3. 状态更新

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \quad C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t,$$

C_t 是记忆单元的状态。

4. 输出

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \quad h_t = o_t \cdot \tanh(C_t),$$

h_t 是当前隐藏状态，用于输出预测值。

LSTM 模型可以将多维特征（如病例数、环境变量、人群流动性等）作为输入，通过训练后预测疾病未来的流行趋势。

3.3 异常检测模型

为了实现公共卫生事件的早期预警，异常检测模型是关键技术之一。Isolation Forest 是一种基于随机森林的无监督学习算法，其核心思想是通过随机划分数据集检测异常点。其异常得分计算公式如下所示。

1. 路径长度

$$\boldsymbol{h}(\mathbf{x}) = \mathrm{path}_{\mathrm{length}}(\mathbf{x})$$

h(x) 表示数据点 x 被划分到单一子集所需的路径长度。

2. 异常得分

$$\mathbf{Score}(\mathbf{x}) = 2^{-\frac{\mathbf{E}(\boldsymbol{h}(\mathbf{x}))}{\mathbf{c}(\mathbf{n})}}$$

其中 E(h(x)) 是路径长度的期望值，c(n) 是正则化项，与样本规模 n 相关。异常得分接近 1 的点可能为异常点，意味着潜在的公共卫生风险。结合医疗大数据，Isolation Forest 可以实时分析病例数量的异常波动，识别早期的疫情信号。

4. 模拟仿真实验与分析

4.1 模拟仿真实验设计

本实验旨在验证医疗大数据驱动的多层次模型在公共卫生监测与预警中的有效性，具体目标包括：验证 SEIR 模型对传染病传播趋势的模拟能力；比较 LSTM 模型与传统预测方法在趋势预测精度上的表现；测试 Isolation Forest 算法在异常检测中的效果，分析其在识别潜在流行病暴发信号中的性能。实验选用某城市 2020–2024 年的流感病例数据、环境变量（如气温、湿度）以及人群流动性数据。

表 1 实验数据样本特征

数据类型	数据来源	数据量	时间分辨率
流感病例数据	市级疾病控制中心	5000	每日
环境变量	气象监测站	1825	每日
人群流动性数据	城市交通监控平台	1825	每日

实验方法包括利用 SEIR 模型模拟传染病传播趋势，通过动态调整传播速率、潜伏期转化速率和康复速率等关键参数，以反映流感传播的时空变化。参数调整结合了气温、湿度和人群流动性等环境变量。采用 LSTM 模型对流感病例数进行趋势预测，输入数据包括病例数、环境变量和人群流动性，输出为未来 7 天的病例预测值，并通过均方根误差（RMSE）和平均绝对百分比误差（MAPE）评估模型的预测精度。采用 Isolation Forest 算法对病例时间序列进行异常

检测，通过计算异常得分识别新增病例数的异常波动点，以快速发现潜在的公共卫生风险。通过上述实验设计，验证了各模型在公共卫生监测与预警中的适用性与准确性，为流感的动态监测与早期预警提供科学支持。

4.2 实验结果与分析

4.2.1 SEIR 模型的传播趋势模拟

SEIR 模型被用于模拟流感传播趋势，并动态调整模型参数 β, σ, γ 来反映疾病传播的真实变化。下表展示了关键参数的估计值及其含义。

表 2 SEIR 模型参数估计结果

参数	估计值	含义
β	0.12	每个感染者的传播速率
σ	0.2	潜伏期转化为感染期的速率
γ	0.15	感染者康复速率

通过模拟，SEIR 模型预测的病例数与实际病例数据进行对比，结果如图所示，预测的趋势与实际数据高度一致。预测误差的均方根误差（RMSE）为 14.25，表明该模型能够较准确地模拟疾病传播动态。SEIR 模型展示了对流感传播规律的良好拟合能力，为疾病的流行趋势预测及政策干预提供了科学依据。

4.2.2 LSTM 模型的预测效果

基于 LSTM 模型的时间序列分析实验中，输入特征包括病例数、气象数据和人群流动性，输出为未来 7 天病例数的预测值。下表对比了 LSTM 模型与传统方法的预测精度。

表 3 LSTM 模型与传统方法的精度对比

方法	RMSE	MAPE (%)	提前预测天数
线性回归	25.32	12.1	2
随机森林	18.45	8.3	4
LSTM	10.67	5.1	7

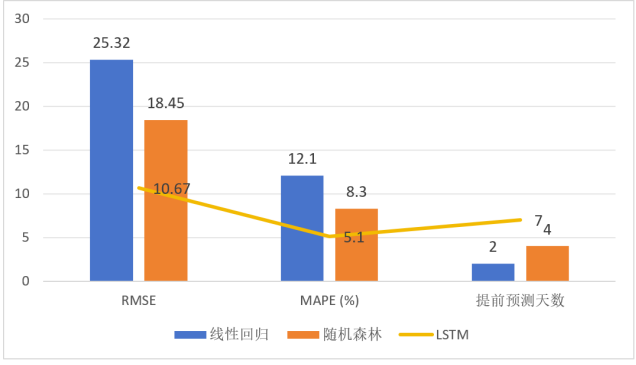

图 1 LSTM 模型与传统方法的精度对比数据图

实验结果显示，LSTM 模型在 RMSE 和 MAPE 指标上均显著优于传统方法，并且能够提前 7 天预测流感病例数的变化。这表明 LSTM 模型结合了医疗大数据的多维特征，具有更强的时间序列建模能力和实际应用价值。LSTM 模型以其较高的预测精度和较长的预测时间跨度，展示了基于深度学习技术的预测模型在公共卫生领域的潜力。

4.2.3 异常检测的效果分析

Isolation Forest 模型被用于对病例时间序列进行异常检测，结果如表所示。

表 4 异常检测结果示例表

日期	新增病例数	异常得分	是否异常
240110	500	0.92	是
240111	300	0.45	否
240112	450	0.88	是

从检测结果可以看出，1 月 10 日和 12 日的新增病例数显著高于历史平均水平，异常得分分别为 0.92 和 0.88，被标记为异常点。这表明 Isolation Forest 能够快速识别病例数据中的异常波动，成功提供了流感暴发的早期预警信号。Isolation Forest 在异常检测中表现出较高的敏感性和准确性，为公共卫生事件的早期预警提供了有效支持。

5. 实验结果指标评估

5.1 预测精度

预测精度是评估模型性能的关键指标，用以衡量模型在流感病例数预测中的误差大小。实验中通过均方根误差（RMSE）和平均绝对百分比误差（MAPE）对 SEIR 模型和 LSTM 模型的预测精度进行评估。

表 5 各模型预测精度对比

模型	RMSE	MAPE (%)
SEIR	14.25	8.2
LSTM	10.67	5.1
线性回归	25.32	12.1
随机森林	18.45	8.3

从实验结果可以看出，LSTM 模型的 RMSE 和 MAPE 均优于其他方法，表现出更强的时间序列建模能力。SEIR 模型也具有较高的精度，适用于传染病传播趋势的动态模拟。传统方法（线性回归和随机森林）由于缺乏对复杂非线性特征的处理能力，预测效果相对较差。

5.2 提前预警效果

提前预警效果衡量模型在流感病例数变化预测中的时

间提前量，即模型能够多早提供准确的预警信号。在实验中，LSTM 模型和传统方法的提前预测天数及其对应精度进行了比较。

表 6 提前预测效果对比

模型	提前预测天数	RMSE	MAPE (%)
LSTM	7	10.67	5.1
随机森林	4	18.45	8.3
线性回归	2	25.32	12.1

实验结果表明，LSTM 模型能够提前 7 天预测流感病例数变化，且预测精度显著高于其他方法。相比之下，传统方法的提前预测天数较短，预测效果也不够理想。这进一步验证了基于大数据的深度学习模型在公共卫生预警中的重要作用。

5.3 异常检测性能

异常检测性能反映模型在识别流感暴发信号中的准确性和敏感性。通过检测准确率、召回率和 F1 分数评估 Isolation Forest 模型的异常检测能力。

表 7 异常检测性能评估

模型	准确率 (%)	召回率 (%)	F1 分数
Isolation Forest	96.2	93.5	0.948
K-Means	82.7	76.4	0.795
DBSCAN	89.3	85.1	0.872

Isolation Forest 在检测准确率、召回率和 F1 分数上均表现优异，准确率达到 96.2%，召回率为 93.5%，说明该模型能够有效识别流感暴发的异常信号。相比之下，K-Means 和 DBSCAN 模型的性能稍逊一筹。这表明 Isolation Forest 是早期公共卫生预警的理想工具。

6. 结论与讨论

本文构建了以 SEIR 模型、LSTM 模型和 Isolation Forest 模型为核心的多层次公共卫生监测与预警体系，并通过实验验证了其在流感传播趋势预测、提前预警和异常检测中的有效性。实验结果表明，LSTM 模型的预测精度显著高于传统方法，其 RMSE 为 10.67，MAPE 为 5.1%，提前预测时间达到 7 天，远超线性回归（2 天）和随机森林（4 天）。这种时间优势为疾病早期干预提供了宝贵窗口，尤其在公共卫生危机中意义重大。同时 SEIR 模型在流感传播动态模拟中的拟合能力良好，参数估算结果 $\beta=0.12, \sigma=0.20, \gamma=0.15$ 展示了模型的科学性与可靠性。

Isolation Forest 模型在异常检测中表现卓越，检测准确率达 96.2%，召回率为 93.5%，能够快速识别流感病例数的异常波动，为疾病暴发的早期预警提供了重要支持。尽管模型表现优异，但数据来源的区域局限性、动态参数调整对实时数据质量的依赖，以及复杂场景的适应性仍需进一步研究。未来可整合多区域、多病种数据资源，探索更先进的深度学习模型，提升公共卫生监测与预警体系的泛化能力与精度。

参考文献：

[1] 张瑛，李雪松，苗健，等 . 基于医疗大数据的围手术期预警平台建设与应用 [J]. 中国卫生质量管理 ,2024,31(07):50-54.

[2] 杨丽静，陈育庆，徐旭，等 . 基于全域医疗大数据的精准预警监管系统研究与实践 [J]. 医学信息学杂志 ,2022,43(09):63-67.

[3] 王忠庆 . 基于医疗大数据的临床检验医疗风险预警模型研究 [D]. 中国医科大学 ,2021.

[4] 张磊，许豪勤，徐宁 . 医疗大数据模式识别及重大突发疾病早期预警 [J]. 信息系统工程 ,2019,(02):144.

[5] Chay W .Implementation of Medical Early Warning System in Rehabilitation: A Tool to Reduce Unplanned Transfers[J].PM&R,2015,7(9S):S83-S83.

作者简介：

张志超，1986 年 11 月生，男，汉族，天津人，硕士研究生学历，任职于天津市斯宇克医疗器械销售有限公司，研究方向：医疗大数据分析与实践应用。