Methods and Applications

Comparison Between Threshold Method and Artificial Intelligence Approaches for Early Warning of Respiratory Infectious Diseases — Weifang City, Shandong Province, China, 2020–2023

Ting Zhang[1,2,3,&]; Liuyang Yang[4,5,&]; Ziliang Fan[6,&]; Xuancheng Hu[1,2,3]; Jiao Yang[1,2,3]; Yan Luo[1,2,3]; Dazhu Huo[7]; Xuya Yu[1,2,3]; Ling Xin[1,2,3]; Xuan Han[1,2,3]; Jie Shan[6]; Zhongjie Li[1,2,3]; Weizhong Yang[1,2,3,#]

ABSTRACT

Introduction: Respiratory infectious diseases, such as influenza and coronavirus disease 2019 (COVID-19), present significant global public health challenges. The emergence of artificial intelligence (AI) and big data offers opportunities to improve traditional disease surveillance and early warning systems.

Methods: The study analyzed data from January 2020 to May 2023, comprising influenza-like illness (ILI) statistics, Baidu index, and clinical data from Weifang. Three methodologies were evaluated: the adaptive dynamic threshold method (ADTM) for dynamic threshold adjustments, the machine learning supervised method (MLSM), and the machine learning unsupervised method (MLUM) utilizing anomaly detection. The comparison focused on sensitivity, specificity, timeliness, and warning consistency.

Results: ADTM issued 37 warnings with a sensitivity of 71% and a specificity of 85%. MLSM generated 35 warnings, with a sensitivity of 82% and a specificity of 87%. MLUM produced 63 warnings with a sensitivity of 100% and specificity of 80%. The initial warnings from ADTM and MLUM preceded those from MLSM by five days. The Kappa coefficient indicated moderate agreement between the methods, with values ranging from 0.52 to 0.62 (P<0.05).

Discussion: The study explores the comparison between traditional methods and two machine learning approaches for early warning systems. It emphasizes the validation of machine learning's reliability and underscores the unique advantages of each method. Furthermore, it stresses the significance of integrating machine learning models with various data sources to enhance public health preparedness and response, alongside acknowledging limitations and the need for broader validation.

Respiratory infectious diseases like seasonal influenza and coronavirus disease 2019 (COVID-19) have the potential to escalate into pandemics or epidemics, rapidly spreading and endangering global public health (*1*). The World Health Organization (WHO) estimates that influenza results in around 1 billion infections, 3–5 million instances of severe illness, and 290,000–650,000 deaths each year (*2*). Timely detection and swift responses to these diseases are crucial in averting outbreaks and controlling the public health threats they bring (*3*).

Threshold-based approaches have traditionally been utilized to promote vigilance regarding respiratory diseases. Models such as the moving percentile method, cumulative sum control chart, and exponentially weighted moving average control chart (*4–5*) evaluate the dynamic nature of time-series data in infectious disease early warning systems. These models issue alerts when reported case numbers meet or exceed predefined thresholds (*6*). With advancements in information technology, there has been a significant shift from reliance on single-source data to incorporating multiple sources. This shift introduces complex analytical processes and the challenge of mitigating noise from large datasets. In the context of COVID-19 management, the application of artificial intelligence (AI) has proven to be exceptionally promising in overcoming these obstacles within surveillance and early warning frameworks (*7*). As a result, the development of robust and dependable AI-driven methods has become crucial in the realm of infectious disease epidemiology.

This study compares the outcomes of traditional methods, specifically the process-credible threshold approach, with two machine learning techniques to assess the suitability and reliability of machine learning methods for early warning systems in infectious disease detection.

METHODS

In this study, "infectious disease early warning" refers to identifying outbreak signs before or during its initial phases through the analysis of infectious disease data from various surveillance sources. Data from January 2020 to May 2023, including influenza-like illness (ILI) statistics, the Baidu index, and clinical data, were analyzed. All methodologies used in this study relied on a uniform and collective data origin.

ILI data from the National Influenza Surveillance Network in China were segmented into China Northern ILI%, Shandong Province ILI%, and Weifang City ILI%. The ILI definition matched the criteria established by the Department of Disease Control and Prevention of the National Health Commission of China, identifying ILI as fever (body temperature ≥38 °C) with cough or sore throat, as referenced (*8*). ILI% represented the ratio of cases among individuals seeking medical care.

The Baidu index, sourced from the publicly accessible Baidu index website, represents the aggregate search frequency of specified keywords on Baidu web pages, with each keyword assigned a particular weight. In the context of "treatment," the index takes into account search terms including "Flu Treatment," "Cold Medicine," "Antipyretic," "Lianhuaqingwen," "What is the most effective flu medicine," "Liuganwan," "Ganmaoqingre," "Banlangen," "Baijiahei," "Oseltamivir," and "Tamiflu." Conversely, the Baidu index for "non-fever symptoms" comprises phrases such as "Fever," "Cough," "Pharyngalgia," "Sore throat," "Runny nose," "Pneumonia," "Chest tightness," "Symptoms of influenza," "Sneezing," "Lacking in strength," and "Muscle soreness" (*9*).

Clinical data from primary and tertiary medical institutions in Weifang City included 21,584,148 chief complaints, 23,128,256 initial diagnoses, 39,486,100 pharmaceutical sales, and 426,171 instances of emergency call data (120). This study focused on respiratory symptoms data, incorporating chief complaints, diagnoses, pharmaceutical sales, and emergency call data (120) with proportional representation.

This study performed a comparative analysis of three early warning methods used in Weifang City, China. The first method improves upon the conventional threshold approach by autonomously determining an optimal threshold, enhancing its practical usability. The second method utilizes supervised machine learning models, whereas the third employs unsupervised machine learning models. Specific details of these models are provided below.

Method 1: Adaptive Dynamic Threshold Method (ADTM)

The ADTM method integrates automatic adjustments into conventional fixed-threshold methods to improve sensitivity and specificity. It consists of five comprehensive phases.

Phase 1. Modeling and parameter setting: Establish models for three distinct scenarios: the beginning of an epidemic season, sudden increases in case numbers, and outliers surpassing historical levels. Each scenario had specific thresholds set through various techniques (Table 1). A total of 1,620 thresholds were determined based on the three warning signal scenarios and different criteria (Figure 1). This process aimed to ensure the model's accuracy in accommodating the dynamic and changing patterns in epidemiological data.

Figure 1 illustrates the criteria for activating alerts in various scenarios. In the epidemic season, an alert is triggered by either "Abrupt Growth" or "Outliers

TABLE 1. Scenarios and criteria for setting early warning thresholds for infectious diseases.

Warning signal scenarios	Criteria
A. Outliers over historical levels	**A1.** Exceeds standard deviations (0.5x, 1x, 1.5x, 2x, 3x) compared to the same period over the last three years, calculated for the past 3 or 7 days. **A2.** Exceeds the 50th to 90th percentiles of case numbers compared to the same period in the past three years, calculated for the past 2 days or weeks.
B. Abrupt Growth	Retrospective time: Two intervals of three days, for a total of six days. **B1.** Absolute change, calculated as the percentage difference between the mean case numbers of two 3-day intervals, with thresholds at 10%, 20%, and 30%. **B2.** Acceleration of absolute change, defined as the difference between absolute changes at adjacent intervals. Acceleration thresholds established at 0.005, 0.01, and 0.015.
C. Epidemic season	Criteria based on exceeding historical data thresholds over 3 or 7 consecutive days. **C1.** Set at 0.5 times the historical mean. **C2.** Standard deviation thresholds at 0.8x, 1x, 1.2x, and 1.5x of the historical average. **C3.** Percentile thresholds at the 50th, 70th, 80th, and 90th percentiles based on historical data.

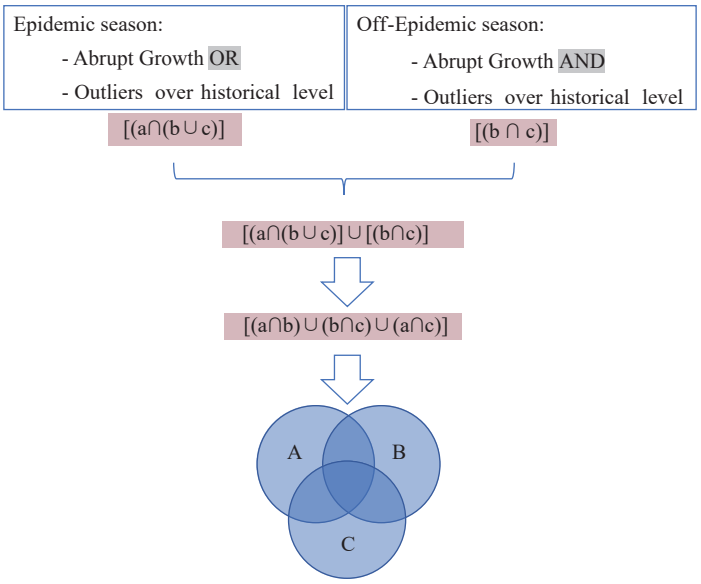


FIGURE 1. Mapping of scenarios and criteria for infectious disease early warning thresholds.

exceeding historical levels." Conversely, during the off-epidemic season, an alert necessitates the concurrent presence of both "Abrupt Growth" and "Outliers exceeding historical levels."

Phase 2. Threshold determination using SoftMax function: This method employs a Softmax function to determine the warning thresholds. The aim is to optimize the balance between timeliness, sensitivity, and specificity for threshold determination, which is crucial for accurate and timely epidemic detection.

Phase 3. Optimal warning strategy for single-source data: Optimally calibrated warning thresholds are applied to single-source data indicators, such as ILI. Warning signals are generated whenever such data surpasses the established threshold, signaling potential health risks that warrant immediate attention.

Phase 4. Integration of multi-source warning signals: Warning signals are synthesized from multi-source warning signals. A comprehensive assessment of warning probability is achieved by calculating a weighted ensemble probability, where each data source is assigned a specific weight. This integration enhances the reliability and accuracy of the warning systems.

Phase 5. Threshold setting for warning probability: A definitive threshold for the warning probability is established to evaluate integrated warning signals. Exceeding this threshold prompts the issuance of an alert, signaling the potential emergence of a public health threat or the initiation of an epidemic.

During these specified phases, the ADTM provides a comprehensive strategy for epidemic surveillance. It incorporates single-source and multi-source data while adjusting thresholds dynamically based on critical epidemiological parameters.

Comparative study of early warning methods: The timeliness of an early warning method was determined by the date of the first warning signal, positioned within the timeline of an outbreak period. The volume of warnings is reflected in the count of days with issued warning signals, as dictated by the warning rules. Consistency was assessed using the Kappa coefficient, which accounts for the probability of random agreement. Statistical significance was attributed to findings with a $P<0.05$.

Method 2: Machine Learning Supervised Method (MLSM)

This approach employs fully supervised learning to reframe the warning issue as a classification task. It accomplishes the categorization of warning levels through the acquisition of multi-source time-series characteristics. The efficacy of early warning for the target metric (Weifang ILI%) is attained by constructing a dataset suitable for supervised learning and utilizing the eXtreme Gradient Boosting (XGBoost) machine learning model (*10*). The XGBoost model, which leverages decision trees and gradient boosting, serves as the underlying framework, which we detail in the Supplementary Materials (available at https://weekly.chinacdc.cn/). Initially, aligning the multi-source time series with the warning labels of the target metric establishes a correspondence between features and labels. Subsequently, these

features and labels are fed into the XGBoost model for training, effectively addressing the supervised learning issue as illustrated in Figure 2.

In the MLSM study, we utilized a training set spanning from January 1, 2020 to November 30, 2022, comprising 1065 days. The test set ranged from December 1, 2022 to May 31, 2023, totaling 182 days. The training set to test set ratio is approximately 6∶1 requirements for dataset partitioning.

Method 3: Machine Learning Unsupervised Method (MLUM)

This approach reconceptualizes the challenge of early warning into a task of anomaly detection. Utilizing unsupervised learning, the model analyzes characteristics of multi-source time series data to identify atypical signals indicative of early warnings. We employ the Isolation Forest algorithm, a machine learning model notably used for its efficacy in anomaly detection (*11*), to thoroughly examine the intrinsic properties of the provided multi-source time series data. The fundamental principle of the Isolation Forest method is that normal and anomalous data points manifest distinct traits; by evaluating and segregating the outliers, the model successfully pinpoints potential anomalies (Supplementary Figure S1, available athttps://weekly.chinacdc.cn/). An advantage of this technique over fully supervised learning is that it eschews the necessity for data labeling, thereby simplifying the implementation of the early warning system (Figure 2). Both the training and test sets in the MLUM were identical to those in the MLSM.

The traditional adaptive dynamic threshold method's effectiveness was rigorously compared with two other methods by assessing their sensitivity and specificity. To establish a reliable benchmark for this evaluation, we used an expert-based consensus. Professionals from the Weifang CDC and senior medical experts reviewed case timelines, labeling moments requiring early warning with a "1" and all other instances with a "0." The application of our technique and the subsequent analyses were performed using Python (version 3.6.13; Python Software Foundation, Fredericksburg, VA, US), aided by the scikit-learn library (version 0.24.2), and the R (version 4.3.1; The R Foundation for Statistical Computing, Vienna, Austria). For Python analyses, the utilized packages included Pandas (1.2.0), Numpy (1.19.5), Xgboost (2.0.3), and Scikit-learn (1.0). For R, the employed packages were ggplot2 (3.4.4), patchwork (1.1.3), scales (1.2.1), dplyr (1.1.4), tidyverse (2.0.0), and readxl (1.4.3).

RESULTS

This study evaluated the performance of the traditional ADTM in comparison with two machine-learning-based methods, MLSM and MLUM, over a period of 182 days from December 1, 2022 to May 31, 2023. ADTM issued 37 warnings with a sensitivity of 71% and a specificity of 85%. MLSM generated 35 warnings with a sensitivity of 82% and a specificity of 87%, while MLUM produced 63 warnings with a sensitivity of 100% and a specificity of 80%. ADTM

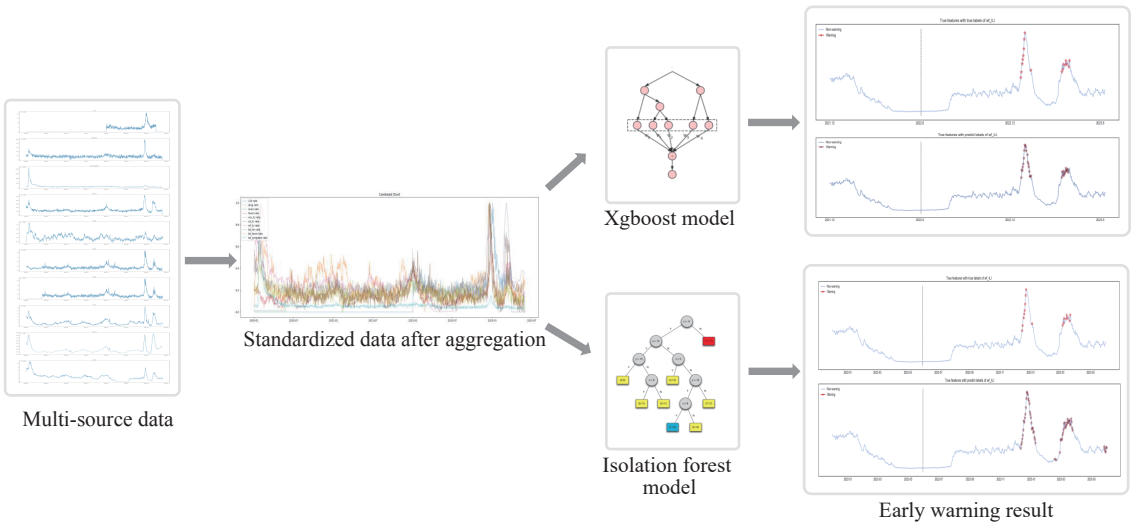


FIGURE 2. Schematic diagram of the MLSM and MLUM models
Abbreviation: MLSM=machine learning supervised method; MLUM=machine learning unsupervised method.

and MLUM issued initial warnings on December 11, with MLSM following on December 16. Pairwise Kappa coefficient analysis indicated significant consistency among these methods (P<0.05) (Figure 3, Table 2).

Panel A illustrates the warning signals derived from the ADTM method for the Weifang ILI% data. Panel B shows the results using the MLSM method, and Panel C depicts the outcomes from the MLUM method. The blue line represents the ILI percentage

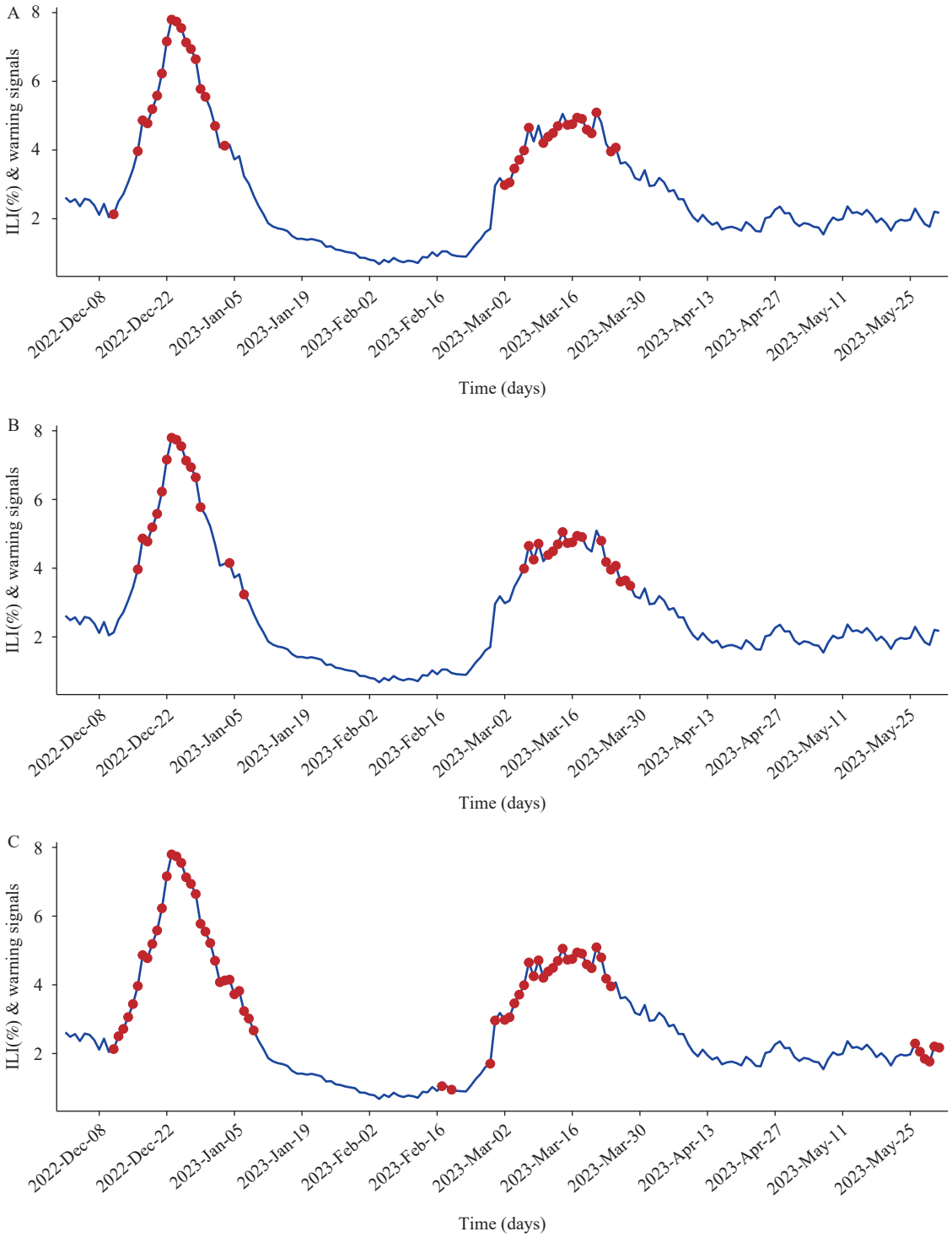


FIGURE 3. Comparative early warning models using three different approaches. (A) ADTM; (B) MLSM; (C) MLUM. Abbreviation: ADTM=adaptive dynamic threshold method; MLSM=machine learning supervised method; MLUM=machine learning unsupervised method.

TABLE 2. Comparative analysis of warning signal detection methods: ADTM, MLSM, and MLUM.

Methods	Warning signals	Timeliness*	Sensitivity (%)	Specificity (%)	Consistency† (Kappa)		
					ADTM	MLSM	MLUM
ADTM	37	Dec 11, 2022	71	85	1.00	§	§
MLSM	35	Dec 16, 2022	82	87	0.62	1.00	§
MLUM	63	Dec 11, 2022	100	80	0.62	0.52	1.00

Note: The values denote the Kappa coefficient.
Abbreviation: ADTM=adaptive dynamic threshold method; MLSM=machine learning supervised method; MLUM=machine learning unsupervised method.
* indicates the date when the warning signals were first initiated.
† represents the consistency, measured by the pairwise Kappa coefficient.
§ signifies a statistical difference with a *P* value of less than 0.05.

curve, while the red points indicate the warning signals.

In addition, in this study, we utilized three evaluation metrics, precision, recall, and F1-score, to evaluate the warning results of the test set, as shown in Supplementary Table S1 (available at https://weekly.chinacdc.cn/).

DISCUSSION

Early warning systems have advanced by utilizing diverse data sources, incorporating big data and machine learning to improve surveillance. This research compares the ADTM approach with MLSM and MLUM, assessing their effectiveness in early warning scenarios. Through establishing ADTM as the reference point, we evaluate the consistency of results from MLSM and MLUM, emphasizing the impact of machine learning on enhancing public health readiness.

This study introduces an enhanced method for infectious disease surveillance, integrating an automatic threshold selection system to enhance adaptability and scalability across various regions. The China Infectious Diseases Automated-Alert and Response System (CIDARS) implements a spatiotemporal early warning model for Type 1 diseases, covering nine infectious diseases, and Type 2 diseases involving 19 infectious diseases, utilizing Fixed-threshold, Temporal, and Spatial detection methods (*10*). However, challenges arose in determining precise thresholds for different regions, times, populations, policies (*11*), and behaviors, hindering rapid adjustments to these factors. To tackle this issue, a dynamic threshold selection function has been developed in this study, enabling real-time adaptation of thresholds, thereby increasing the method's versatility and facilitating its application across diverse geographical areas.

The three methodologies, ADTM, MLSM, and MLUM, exhibit varying effectiveness and suitability in early warning systems. MLSM and MLUM represent the fundamental paradigms of machine learning, each offering unique approaches to problem solving. ADTM and MLUM are particularly relevant for timely anomaly detection crucial for outbreak response. ADTM excels in specificity, reducing false alarms and saving resources, but may lack sensitivity in detecting certain anomalies. MLSM strikes a balance between sensitivity and specificity, albeit with less straightforward interpretability. MLUM stands out for its high sensitivity, benefiting disease detection at the expense of specificity, making it valuable for conditions with significant clinical impacts. Validating the reliability of machine learning methods in infectious disease early warning, the study uses ADTM as a benchmark. Machine learning's computational strength, combined with independence from traditional benchmarks, bears promise for future applications. However, caution is advised as predictive models, with moderate Kappa coefficient agreement, are not infallible and should not be the sole determinants of public health decisions. A more robust approach involves integrating diverse data sources and surveillance methods with predictive models to enhance early warning system reliability and effectiveness, mitigating the limitations of single-model predictions and fortifying public health strategies.

The methodology of the study has limitations, particularly in finding a dependable benchmark for early warning models, notably with machine learning. The study aimed to compare models under similar conditions without in-depth exploration of their intricacies. The MLUM model prioritized timeliness and sensitivity, albeit with a trade-off in specificity due to its parameter settings. Future research may consider more sophisticated models to enhance accuracy. Furthermore, the study was confined to the Weifang area, suggesting the necessity for broader validation in

other regions in subsequent work.

Conflicts of interest: No conflicts of interest.

Funding: Supported by the CAMS Innovation Fund for Medical Sciences (2021-I2M-1-044, 2023-I2M-3-011) and the National Key Research and Development Program of China (2023YFC2308701).

doi: 10.46234/ccdcw2024.119

[#] Corresponding author: Weizhong Yang, yangweizhong@cams.cn.

[1] School of Population Medicine and Public Health, Chinese Academy of Medical Sciences (CAMS) & Peking Union Medical College (PUMC), Beijing, China; [2] State Key Laboratory of Respiratory Health and Multimorbidity, Beijing, China; [3] Key Laboratory of Pathogen Infection Prevention and Control (Peking Union Medical College), Ministry of Education, Beijing, China; [4] The Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Kunming City, Yunnan Province, China; [5] School of Data Science, Fudan University, Shanghai, China; [6] Weifang Center for Disease Control and Prevention, Weifang City, Shandong Province, China; [7] School of Health Policy and Management, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China.

[&] Joint first authors.

Submitted: March 05, 2024; Accepted: June 24, 2024

REFERENCES

1. Van Kerkhove MD, Ryan MJ, Ghebreyesus TA. Preparing for "Disease X". Science 2021;374(6566):377. https://doi.org/10.1126/science.abm7796.
2. Krammer F, Smith GJD, Fouchier RAM, Peiris M, Kedzierska K, Doherty PC, et al. Influenza. Nat Rev Dis Primers 2018;4(1):3. https://doi.org/10.1038/s41572-018-0002-y.
3. Hamalaw SA, Bayati AH, Babakir-Mina M, Benvenuto D, Fabris S, Guarino M, et al. Assessment of core and support functions of the communicable disease surveillance system in the Kurdistan Region of Iraq. J Med Virol 2022;94(2):469 – 79. https://doi.org/10.1002/jmv.27288.
4. Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). J Urban Health 2003;80(2 Suppl 1):i89-96. http://dx.doi.org/10.1007/pl00022319.
5. Li ZJ, Lai SJ, Zhang HL, Wang LP, Zhou DL, Liu JZ, et al. Hand, foot and mouth disease in China: evaluating an automated system for the detection of outbreaks. Bull World Health Organ 2014;92(9):656 – 63. https://doi.org/10.2471/BLT.13.130666.
6. Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. J R Stat Soc Ser A: Stat Soc 2012;175(1):49 – 82. https://doi.org/10.1111/j.1467-985X.2011.00714.x.
7. MacIntyre CR, Chen X, Kunasekaran M, Quigley A, Lim S, Stone H, et al. Artificial intelligence in public health: the potential of epidemic early warning systems. J Int Med Res 2023;51(3):3000605231159335. http://dx.doi.org/10.1177/03000605231159335.
8. Feng LZ, Zhang T, Wang Q, Xie YR, Peng ZB, Zheng JD, et al. Impact of COVID-19 outbreaks and interventions on influenza in China and the United States. Nat Commun 2021;12(1):3249. https://doi.org/10.1038/s41467-021-23440-1.
9. Yang LY, Zhang T, Han X, Yang J, Sun YX, Ma LB, et al. Influenza epidemic trend surveillance and prediction based on search engine data: deep learning model study. J Med Internet Res 2023;25:e45085. https://doi.org/10.2196/45085.
10. Yang WZ, Li ZJ, Lan YJ, Wang JF, Ma JQ, Jin LM, et al. A nationwide web-based automated system for outbreak early detection and rapid response in China. Western Pac Surveill Response J 2011;2(1):10 – 5. https://doi.org/10.5365/WPSAR.2010.1.1.009.
11. Han SS, Zhang T, Lyu Y, Lai SJ, Dai PX, Zheng JD, et al. The incoming influenza season - China, the United Kingdom, and the United States, 2021-2022. China CDC Wkly 2021;3(49):1039 – 45. https://doi.org/10.46234/ccdcw2021.253.

SUPPLEMENTARY MATERIAL

Model Introduction

XGBoost is a popular ensemble learning algorithm called eXtreme Gradient Boosting, commonly applied for classification and regression purposes. It utilizes decision trees and the gradient boosting technique to enhance model performance through iteratively training new decision trees.

XGBoost relies on decision trees as its fundamental components, with each tree functioning as a weak learner. Decision trees encompass nodes, branches, and leaves, where nodes split based on features and leaf nodes correspond to output values. The model's goal is to minimize an objective function consisting of a loss function and regularization term. This objective function evaluates the model's performance, aiming to minimize it by seeking new decision trees in each iteration. This optimization process can be described by Equation 1, which also represents the objective function. In Equation 1, y_t denotes the objective function, l represents the loss function, y_i is the actual label, $\hat{y}_i$ stands for the model's predicted label, K denotes the number of trees, and $\Omega(f_k)$ is the regularization term.

$$y_t = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

XGBoost employs the gradient boosting strategy to minimize the objective function gradient at each iteration.

The process of generating a new tree includes fitting the negative gradient of the current model. This guides the creation of new trees to prioritize poorly-performing samples from the previous model.

The negative gradient can be represented by Equation 2.

$$G_n = -\frac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i} \quad (2)$$

To address overfitting, XGBoost incorporates regularization methods such as weight decay (L2 regularization) and minimum split loss to manage tree depth and leaf node weights. The calculation for the regularization term is detailed in Equation 3.

$$\Omega(f_k) = \frac{1}{2}\sum_{j=1}^{L} w_j^2 \quad (3)$$

The ultimate prediction is calculated by summing the predicted values of all generated trees, with each tree's impact adjusted by the learning rate. This methodology enables XGBoost to boost performance by amalgamating diverse decision trees.

XGBoost is a potent machine learning algorithm, well-suited for medium to large-scale datasets and intricate classification tasks. Optimal performance is attainable via meticulous parameter adjustment. Leveraging multithreading and parallel computing, the model demonstrates efficient performance, enabling swift training on extensive datasets. Incorporating regularization terms aids in averting overfitting and enhancing the model's generalization capabilities. XGBoost offers insights on feature importance, facilitating comprehension of the model's sensitivity to specific features. It adeptly manages missing values without necessitating supplementary processing. Additionally, the algorithm accommodates diverse loss functions and evaluation metrics, rendering it versatile across various problem types.

In this study, the model successfully meets the demands of the research task by providing scientifically precise predictions for influenza-like case activity levels.

The Isolation Forest

Forest algorithm model was first introduced by Fei Tony Liu et al. It focuses on anomaly detection by differentiating typical data from anomalies, enabling efficient classification by isolating the anomalous data points.

The primary operational concept of the Isolation Forest model involves randomly selecting a feature from the dataset and choosing a separation value within its range. Samples are then split into branches based on this value. This recursive process continues for each branch until only one sample is left or the defined recursion depth is met.

Anomalous data is distinguished from normal data by its unique features, requiring fewer partitioning steps for isolation. Conversely, normal data necessitates more steps for isolation, leading to a longer path to the endpoint. The model assesses the path length for each data point to reach isolation, with a shorter path indicating early isolation and suggesting a higher likelihood of anomaly.

To provide a clearer understanding of the Isolation Forest model's training methodology, a straightforward example is discussed. Refer to Supplementary Figure S1, which presents a set of data points plotted along a number line, with their values arranged in ascending order. The objective is to identify any anomalies within these data points. The initial step involves determining the median of all values, which lies between the maximum and minimum data points, to serve as the initial split value. Following this initial split, data point Z becomes separated. Subsequently, the median value amongst the remaining data points, A through I, is selected to define the second split, consequently isolating data points F through I and leaving data points A through E. This bifurcation proceeds recursively until each data point stands alone. Ultimately, data point Z is sequestered in just one split, whereas data point E requires four splits for isolation. The fewer partitions required to isolate a data point signal a higher likelihood of it being anomalous, hence data point Z is deemed more likely to be an outlier compared to data point E.

Evaluation of Model Warning Results

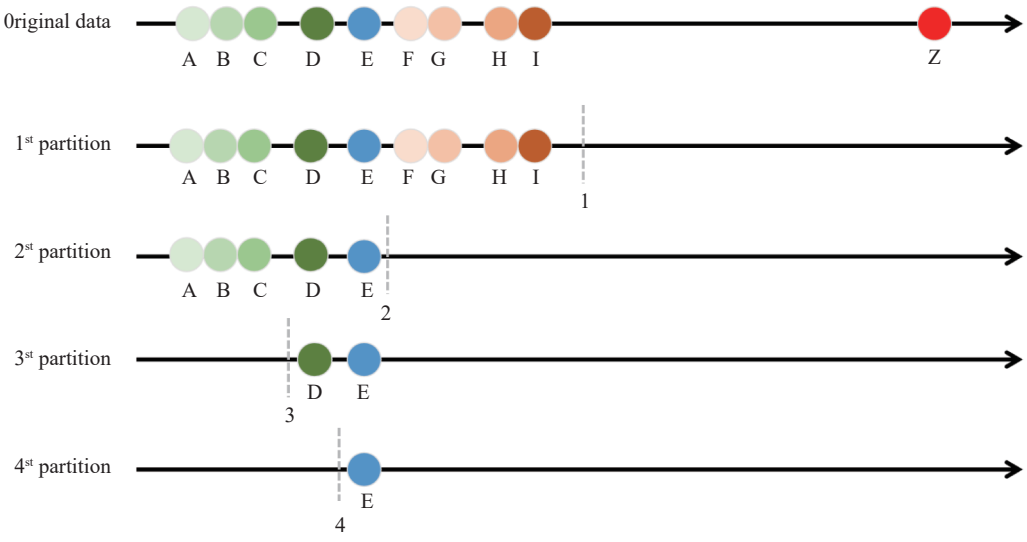
In this study, three evaluation metrics, Precision, Recall, and F1-Score, were utilized to assess the warning outcomes of the test set. The calculation equations for these metrics are presented below:

$$\text{Precision (Pre)} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall (Rec)} = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2 \times \frac{Pre \times Rec}{Pre + Rec} \tag{6}$$

True positive (TP) refers to the count of normal events correctly classified as normal, while true negative (TN) denotes the count of abnormal events accurately identified as abnormal. False positive (FP) indicates the number of abnormal events erroneously detected as normal, and false negative (FN) represents the count of normal events mistakenly presented as abnormal. The detailed evaluation results are presented in Supplementary Table S1.



SUPPLEMENTARY FIGURE S1. Training process of the Isolation Forest model.

SUPPLEMENTARY TABLE S1. Evaluation results based on the XGBoost warning model.

Weighted average	Precision	Recall	F1-Score
XGBoost	0.93	0.88	0.90
Isolation forest	0.95	0.82	0.86