BioData Mining

RESEARCH

Open Access

From COVID-19 to monkeypox: a novel predictive model for emerging infectious diseases

Deren Xu[1*], Weng Howe Chan[2*], Habibollah Haron[1], Hui Wen Nies[1] and Kohbalan Moorthy[3]

*Correspondence:
2008xuderen@gmail.com;
cwenghowe@utm.my

[1] Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[2] UTM Big Data Centre, Ibnu Sina Institute For Scientific and Industrial Resarch Universiti Teknologi, Johor Bahru 81310, Malaysia
[3] Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang 26600, Malaysia

Abstract

The outbreak of emerging infectious diseases poses significant challenges to global public health. Accurate early forecasting is crucial for effective resource allocation and emergency response planning. This study aims to develop a comprehensive predictive model for emerging infectious diseases, integrating the blending framework, transfer learning, incremental learning, and the biological feature Rt to increase prediction accuracy and practicality. By transferring features from a COVID-19 dataset to a monkeypox dataset and introducing dynamically updated incremental learning techniques, the model's predictive capability in data-scarce scenarios was significantly improved. The research findings demonstrate that the blending framework performs exceptionally well in short-term (7-day) predictions. Furthermore, the combination of transfer learning and incremental learning techniques significantly enhanced the adaptability and precision, with a 91.41% improvement in the RMSE and an 89.13% improvement in the MAE. In particular, the inclusion of the Rt feature enabled the model to more accurately reflect the dynamics of disease spread, further improving the RMSE by 1.91% and the MAE by 2.17%. This study underscores the significant application potential of multimodel fusion and real-time data updates in infectious disease prediction, offering new theoretical perspectives and technical support. This research not only enriches the theoretical foundation of infectious disease prediction models but also provides reliable technical support for public health emergency responses. Future research should continue to explore integrating data from multiple sources and enhancing model generalization capabilities to further enhance the practicality and reliability of predictive tools.

Keywords: Emerging infectious disease prediction, Transfer learning, Incremental learning, Biological feature Rt, Blending framework

Introduction

Emerging infectious diseases present a significant challenge to global public health [30, 48, 51]. In recent years, accelerated globalization and ecological changes have led to a marked increase in the rate and extent of infectious disease transmission [39, 40]. This surge not only places immense pressure on health systems but also threatens global

economic and social stability [23]. Emerging infectious diseases, such as COVID-19 and monkeypox, have triggered widespread health crises in a relatively short timeframe, and their unpredictability often renders traditional response strategies ineffective [10].

Early prediction is critical for an effective public health response [1]. Timely and accurate outbreak forecasts not only assist decision-makers in allocating resources proactively but also facilitate the implementation of more effective control measures, thereby minimizing the social and economic impacts of outbreaks [71, 72]. Specifically, early prediction plays a vital role in supporting vaccination strategies, healthcare resource allocation, and public health interventions. To achieve this, outbreak forecasting must rely on quantitative analyses of infectious disease transmission trends and consider complex factors such as pathogen biology, modes of transmission, and population movement. Consequently, the development of efficient and reliable forecasting tools—particularly those that deliver accurate predictions in the early stages of emerging infectious diseases—represents a significant challenge in public health [62].

Despite some progress in predicting early emerging infectious diseases, significant limitations persist [55, 71, 72]. Traditional prediction models frequently produce delayed results due to the untimeliness of data when addressing rapidly evolving epidemic situations. Furthermore, fixed-interval time series data struggle to capture the dynamic nature of disease transmission accurately [86]. Existing machine learning methods often rely on single-point models that overlook the influence of factors such as human mobility, regional variations, and epidemiological knowledge of disease transmission [42, 44]. Furthermore, the underutilization of real-time data and dependence on small-scale datasets severely compromise the accuracy of early emerging infectious disease prediction models [2]. Although internet and search engine data demonstrate significant potential for early prediction, effectively utilizing these unstructured data remains a challenge [28, 29, 68, 69, 73]. Models such as SIR and SEIR, while useful for early predictions, tend to underestimate the resources required owing to data limitations and evolving outbreak characteristics. This underscores the necessity of dynamically updating model input data [43, 42, 44]. Therefore, the integration of epidemiological knowledge, real-time data, and dynamic model updates has become essential for enhancing the accuracy and effectiveness of early emerging infectious disease prediction.

The primary objective of this study is to increaes the accuracy and utility of early predictions of emerging infectious diseases by integrating existing multimodal approaches, utilizing transfer learning and incremental learning techniques, and introducing biometric Rt. We aim to address the limitations of current models regarding data scarcity and predictive generalizability. By performing feature migration from COVID-19 data and dynamically updating real-time information, we seek to improve the model's performance in predicting emerging infectious diseases, including monkeypox.

The contributions of this study are primarily in the following areas. First, with the acceleration of globalization and ecological changes, the frequency and spread of emerging infectious diseases are increasing, posing significant challenges to global public health [30, 45, 51]. Accurate early prediction can provide critical decision support to public health authorities, facilitating the effective allocation of resources and the development of emergency response measures, thereby reducing the spread and impact of outbreaks [7, 24, 60]. Second, this study explored the potential of multimodel fusion

to provide more stable and reliable predictions by integrating the strengths of multiple models [26, 68, 69, 73]. By introducing biometric Rt, the models are not only able to capture disease transmission trends, but also able to evaluate the effectiveness of prevention and control measures, thus providing a basis for developing more scientific public health strategies. In addition, the application of transfer learning and incremental learning in the context of sparse and continuously updated data offers unique advantages to help address emerging infectious diseases [19, 38]. By drawing on knowledge from previous outbreaks and updating data in real time, these approaches significantly improve the predictive accuracy and adaptability of models, which is crucial for rapid response in the early stages of an outbreak, helping to implement timely control measures to slow down the spread of the disease.

Related work

Traditional infectious disease prediction models, such as the susceptible-infected-recovered (SIR) and susceptible-exposed-infected-recovered (SEIR) models, have been widely utilized for modelling and forecasting the spread of diseases [67]. These models predict disease transmission trajectories by delineating various states of the population and relying on epidemiological parameters. However, models such as SIR and SEIR have significant limitations. They depend on rigid assumptions and are unable to effectively account for complex social behaviors and population movements [32].

With the advancement of computing technology, the application of machine learning in predicting infectious diseases has steadily increased. In recent years, techniques such as neural networks, decision trees, and support vector machines have been utilized for the early prediction of infectious diseases [70]. However, these methods still face limitations in capturing the complex and dynamic characteristics of diseases [68, 69, 73]. Existing machine learning approaches encounter several challenges: First, most models rely heavily on historical data, which leads to reduced prediction accuracy when confronted with limited data or emerging diseases [34, 77, 78, 80, 79, 81]. Second, these models struggle to operate effectively with real-time data updates and cannot adapt to the evolving dynamics of epidemics [47]. This issue is particularly pronounced in the prediction of emerging infectious diseases such as COVID-19 and monkeypox.

Internet big data, including social media, search engines, and news data, have become important resources for predicting infectious diseases in recent years. The real-time nature and extensive coverage of these data provide significant potential for early warning [77, 78, 80]. For example, by analysing COVID-19-related web search behaviors, researchers were able to detect early signals of disease transmission [28, 29]. However, extracting effective information from massive amounts of unstructured data remains a technical challenge, particularly regarding automatic keyword filtering and data cleaning [14]. Additionally, geographical differences and delays in information dissemination within internet data limit its applicability on a transnational or global scale [16].

Early prediction of emerging infectious diseases Against the problem of data scarcity, migration learning allows the model to learn from similar disease data, thus improving the generalization of the prediction of emerging diseases [57]. On the other hand, incremental learning improves the timeliness of prediction by enabling the model to be dynamically updated as new data arrives without the need to retrain the entire model

[88]. In addition, multimodel fusion has been an effective method for improving the stability and accuracy of prediction models in recent years [77, 78, 80]. By integrating the advantages of different models, more robust prediction results can be obtained. The introduction of biological features such as the effective reproduction number (Rt) also enhances the prediction ability of the model. These techniques show great potential for the early prediction of emerging infectious diseases.

Method

Data collection and processing

This study utilized the Mexican COVID-19 dataset and the US Monkeypox dataset provided by Our World in Data [20]. The COVID-19 dataset covers the period from April 1, 2020, to March 31, 2023, and includes various key indicators such as daily new cases (new_cases), total confirmed cases (total_cases), daily new deaths (new_deaths), total deaths (total_deaths), smoothed data, and rates calculated per million people. These datasets are widely used in epidemiological research because of their completeness and accuracy [36]. The Monkeypox dataset spans from May 10, 2022, to February 27, 2024, encompassing key metrics such as daily new cases, 7-day averages, and total cases. Given the presence of missing values in the original dataset, a median imputation method was used to fill these gaps. Owing to the unique characteristics of infectious disease time-series data, outliers were retained to preserve the integrity and authenticity of the epidemiological trends.

Construction of a blending model based on COVID-19

Covid-19 feature engineering

Feature engineering constitutes a critical phase in data processing, particularly when predicting the spread of diseases. Well-designed features can significantly enhance a model's predictive capabilities [11]. In the present study, a series of feature engineering techniques were applied to the Mexican COVID-19 dataset to capture dynamic changes and transmission trends of the disease. To comprehend temporal trends in COVID-19 case numbers, 7-day and 14-day moving averages were computed. These features assist the model in recognizing recent trends in the case of increases or decreases.

Lag features, a common technique in time series analysis, enable the model to capture the autocorrelation inherent in sequential data [5]. Lagged features for new cases were generated with 1-day and 7-day intervals. Differential features were also employed to capture the rate of change in the series, which is crucial for detecting accelerations or decelerations in disease transmission [31]. Additionally, growth rate features were created to provide an alternative perspective on the relative changes in case numbers by calculating the daily percentage increase in new cases [54].

The time point features, including the year, month, and day of the week, were extracted from the data. These features assist the model in identifying potential seasonal patterns or weekly cyclical changes. Owing to the potential generation of missing values (NA) from the computation of sliding windows and lag features, rows containing NA were removed after completing feature engineering to ensure data integrity for model training and testing. Through the implementation of the aforementioned feature engineering processes, a comprehensive and insightful set of features was prepared for the

COVID-19 case prediction model. These features not only enhance the data's expressiveness but also improve the predictive model's accuracy and interpretability.

Feature selection is a critical aspect of constructing effective predictive models, aiming to identify features that most significantly impact the target variable—in this case, the number of new COVID-19 cases ("new_cases") [58]. This study employs the XGBoost regression model for feature selection. XGBoost (extreme gradient boosting) is a widely used gradient boosting decision tree (GBDT) algorithm that provides importance scores for each feature during the training process. These scores are based on each feature's contribution to the model's predictive performance, such as reducing errors in the training data. XGBoost uses regularization techniques during the construction of decision trees to prevent overfitting, leading to the selection of fewer but more informative features. By efficiently leveraging these features, it helps identify those with the greatest predictive power for the target variable [64].

During training, certain features may not be selected for any splits in the trees, indicating that they do not significantly contribute to the model's predictive ability. Observing these overlooked features allows for their removal to simplify the model (as illustrated in Table 1). Features with zero contribution values were excluded. This method precisely selects useful features, thereby optimizing the model's performance [74]. These insights can guide future research to more accurately predict and manage the spread of COVID-19 and similar infectious diseases.

In the data splitting phase, the procedure was conducted in two steps. Initially, 60% of the data were allocated for model training to ensure sufficient data for this purpose.

Table 1 Contribution of features assessed via the XGBoost feature importance assessment feature

Feature name	Contribution value
new_cases_per_million	0.9995
new_deaths	6.1137
new_cases_lag1	5.2678
day_of_week	4.8319
new_cases_diff	4.6435
new_cases_smoothed	4.5880
new_deaths_smoothed	4.1125
new_cases_14d_avg	3.7031
new_cases_lag7	3.4105
total_cases	3.2471
month	2.6358
new_cases_growth_rate	2.5328
total_deaths	0.0
total_cases_per_million	0.0
new_cases_smoothed_per_million	0.0
total_deaths_per_million	0.0
new_deaths_per_million	0.0
new_deaths_smoothed_per_million	0.0
new_cases_log	0.0
new_cases_7d_avg	0.0
year	0.0

The remaining 40% of the dataset was then equally divided into validation and test sets. To maintain the chronological order inherent in time series data, shuffling was disabled by setting `shuffle = False`, thereby preventing random shuffling of the data [12].

Model selection

In a previous empirical study, the effectiveness of various machine learning and deep learning models in predicting outcomes from the COVID-19 dataset was compared. These models demonstrated strong performance in time series prediction. The study employed the NSGA-II algorithm to evaluate the models on the basis of prediction accuracy, generalization ability, and computational efficiency. Models such as ridge regression, decision trees (DT), and XGBoost, which are selected via the multiobjective optimization method NSGA-II, exhibit commendable performance in terms of accuracy, generalizability, and computational efficiency [76]. Consequently, ridge regression, decision trees (DT), and XGBoost were chosen as the base models for blending, as they did not yield significant differences in prediction results. Linear regression was utilized as the metamodel, considering the generalization ability and computational efficiency of the models.

Blending ensemble model

This study proposes a blended ensemble model that integrates three distinct base models: ridge regression, decision tree regressor, and XGBoost regressor. Each of these models employs different underlying algorithms, enabling the capture of unique data characteristics from multiple perspectives. This diversity in model architecture forms the foundation for developing a robust and flexible blending ensemble learning framework (see Fig. 1).

Ridge regression Ridge regression is a linear regression model that prevents overfitting by adding an L2 regularization term to the loss function [27]. The L2 regularization term is the sum of the squared parameters. Given that the blending is trained on the COVID-19 dataset, it is important to note that there are significant differences in the data distributions and ranges between the monkeypox and COVID-19 datasets. FastICA is highly sensitive to the data distribution and range [9]. Therefore, the ridge model trained on the COVID-19 dataset via FastICA is replaced with principal component analysis (PCA). Setting alpha = 0.1 ensures that the model parameters do not become too large, thereby enhancing the model's generalization ability. In epidemiological forecasting, data features may exhibit multicollinearity, which ridge regression can effectively handle [35]. Furthermore, it can balance the bias and variance of the model by adjusting the regularization parameter, providing stable and reliable predictive outcomes.

Decision tree regression The decision tree is a nonparametric supervised learning method applicable to regression and classification tasks [66]. It constructs a tree model by recursively partitioning the dataset into smaller subsets, where each node represents a feature, each branch represents a possible value of that feature, and each leaf node

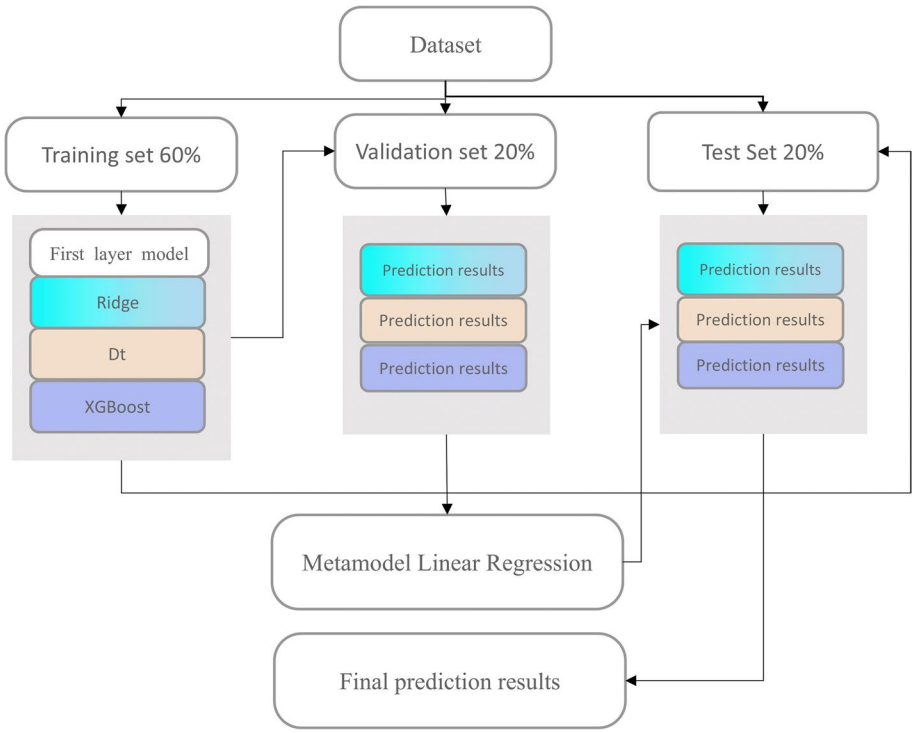


Fig. 1 Blending model structure diagram

represents a prediction. Decision trees can capture complex nonlinear relationships and are easily interpretable [84]. They perform well with high-dimensional data and missing values. Owing to their simple structure, they can be trained and predicted quickly, making them highly suitable for real-time prediction tasks.

XGBoost XGBoost (extreme gradient boosting) represents an advanced version of boosting trees, enhancing predictive performance by incrementally adding new trees [53]. Combining weighted linear models and tree models, it effectively handles nonlinear relationships and demonstrates strong generalization capabilities. Renowned for its efficient computation and robust predictive power, XGBoost excels particularly in managing large-scale data. Because it is capable of automatically addressing missing values and conducting feature selection, it is well suited for intricate time series forecasting tasks [4].

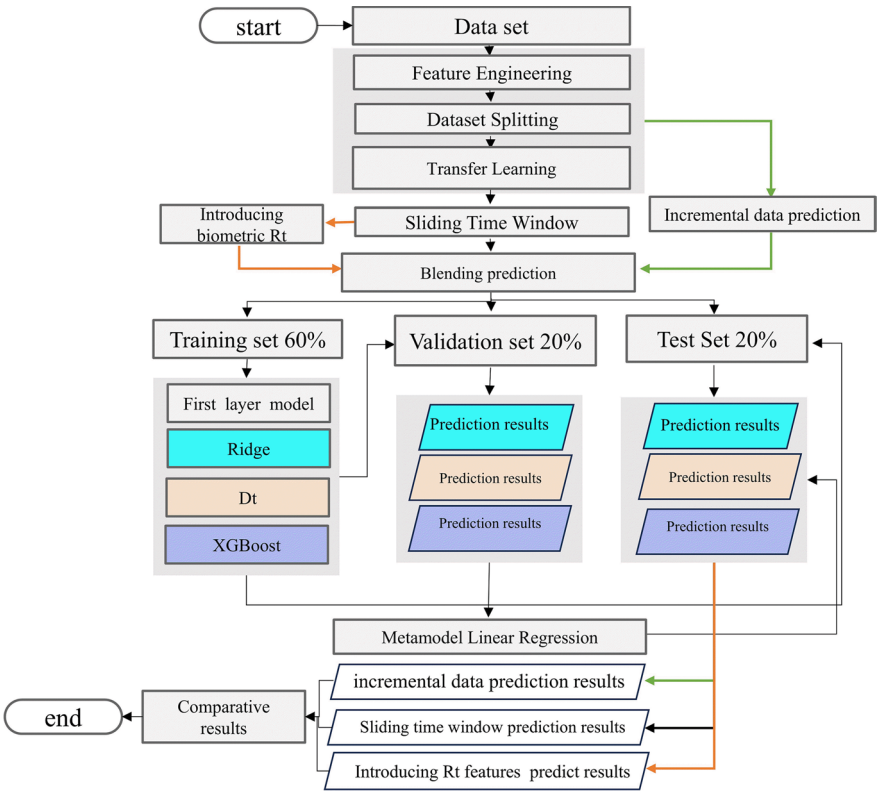
Linear regression Linear regression serves as a fundamental regression technique that minimizes the error between the predicted and actual values by fitting a linear equation [82]. It assumes a linear relationship between the target and input variables, estimating model parameters through minimizing the mean square error. The choice of linear regression as a meta-model is attributed to its simplicity, interpretability, and efficiency in fitting and predicting. Within the blending framework, linear regression effectively amalgamates predictions from various base models to enhance the overall predictive

performance. Given that base models capture intricate data relationships, the meta-model merely requires a linear combination of base model predictions.

In the evaluation of model performance metrics, the hybrid model demonstrates strong generalizability and accuracy on the COVID-19 dataset from Iran, Indonesia, and Chile. These results underscore the hybrid approach's capacity to minimize errors and enhance stability by integrating predictions from various models (see Table 2).

This blending base model and meta-model are saved as a pickle file for the early prediction of emerging infectious diseases.

Modelling enhancements for the early prediction of emerging infectious diseases



Methodology flowchart of the enhanced blended ensemble model

Monkepox feature engineering

Introducing additional dimensions of features can enhance the model's understanding of disease transmission patterns and improve its predictive ability [41, 85]. Given the time-dependent nature of infectious disease spread, incorporating lag features and smoothing features can capture the dynamic changes in time series data. The specific features are as follows:

The extraction of Year, Month, and Day of the Week from dates is typically performed via date-time libraries in programming languages such as Python's datetime module. This process aids in capturing the effects of seasonal and periodic variations in models

Table 2 Comparison of blending and its base model performance under different metrics

Model Name	Accuracy (RMSE)	Generalization (RMSE,Iran)	Generalization (RMSE,Indonesia)	Generalization (RMSE,Chile)	Model Training Time
Ridge	11.52	114.85	16.25	23.24	0.03 s
XGBoost	33.07	13.06	13.51	7.98	0.36 s
Decision Tree	24.48	8.14	6.73	4.44	0.07 s
Blending	1.87	1.95	1.67	2.07	0.40 s

[21, 56]. New_Cases_14d_Avg is calculated as the average number of new cases in the past 14 days, achieved through the application of rolling window functions. This feature helps smooth daily fluctuations, offering a clearer view of disease transmission trends. New_Cases_Cubic represents the cube of the number of new cases, potentially assisting models in identifying nonlinear patterns in case growth. New_Cases_Diff indicates the daily increment in new cases compared with the previous day, reflecting the immediate changes in disease transmission speed. The New_Cases_Growth_Rate is calculated as the ratio of new cases on a given day to those on the previous day, providing insight into the growth rate of new cases. New_Cases_Lag1 and New_Cases_Lag7 denote the number of new cases one day and seven days prior, respectively. These lag features help models understand the short-term and medium-term dynamics of disease transmission. New_Cases_Per_Million represents the number of new cases per million people, standardizing case numbers for comparison across different regions. New_Cases_Smoothed refers to new cases processed by moving averages or other smoothing techniques to reduce the impact of daily fluctuations [6]. Week_of_Year indicates the week number within a year, aiding models in capturing the periodic effects of weeks.

Dataset splitting

The entire dataset was divided into a training set and a validation set, with the training set comprising 60% and the validation set comprising 40%. In this step, we set shuffle = False to preserve the temporal order of the data, as time series data exhibit temporal dependencies, and shuffling the order could disrupt the model's ability to learn these temporal features accurately. The validation set was subsequently further divided into a validation set and a test set, each accounting for 20%. Similarly, shuffle = False was employed to maintain the temporal sequence, ensuring that the model has an adequate amount of data for training.

Transfer learning

This study, employs transfer learning to address the challenge of limited data in the early stages of the monkeypox epidemic. Transfer learning leverages existing knowledge to address new but related problems, significantly enhancing model performance on novel tasks [37]. Specifically, we utilize features from the COVID-19 dataset for feature transfer and adapt the model to suit the monkeypox data. Feature transfer, an essential method in transfer learning, aligns features from the source domain (COVID-19) with the target domain (monkeypox), enabling models trained in the source domain to be

applied to the target domain. The following outlines the specific steps involved in our feature transfer process:

Feature alignment

In the COVID-19 dataset, features such as 'new_cases_per_million', 'new_deaths', 'new_cases_lag1', 'day_of_week', 'new_cases_diff', 'new_cases_smoothed', 'new_deaths_smoothed', 'new_cases_14d_avg', 'new_cases_lag7', 'total_cases', 'month', and 'new_cases_growth_rate' have been identified as valuable for training the model as they contribute to the target variable. Similarly, in the monkepox dataset, features such as 'new_cases_cubic', 'new_cases_diff: 7-Day Average', 'new_cases_lag1', 'new_cases_lag7', 'day_of_week', 'total_cases', 'new_cases_14d_avg', 'new_cases_growth_rate', 'week_of_year', 'month', and 'year' are considered valuable for predicting the target variable. Retaining the relevant features from each dataset and aligning them for analysis is crucial. Notably, the COVID-19 dataset includes unique features such as 'new_cases_per_million', 'new_deaths', 'new_cases_smoothed', and 'new_deaths_smoothed'. To ensure full alignment of the features between the two datasets, it is necessary to create missing features in each dataset. For the COVID-19 dataset: new_cases_cubic can be derived by calculating the cube of new_cases. week_of_year can be derived from the date, and Year can also be derived from the date. For the monkepox dataset: new_cases_per_million can be calculated using a population of 335.9 million people. Regarding new_deaths, a simple assumption is made that deaths constitute a fixed percentage of new cases. In this study, a 2% mortality rate is assumed, meaning that for every 100 new cases, there are expected to be 2 deaths. This ratio is an estimate, as early stages of novel infectious diseases often lack relevant information, typically on the basis of data from similar outbreaks. new_cases_smoothed can be created by computing the moving average of new_cases. Similarly, new_deaths_smoothed is calculated by applying a moving average to new_deaths to smooth daily fluctuations and provide a more stable trend of death cases. In this study, a 7-day moving average is employed. This methodology is commonly used for time series data to help reveal long-term trends and reduce the impact of short-term fluctuations. Figure 2. depicts the feature after completing feature alignment. Owing to the limited availability of early monkeypox data, the feature-aligned datasets for COVID-19 and monkeypox were merged into a new dataset.

Feature transfer

The pretrained blending model, which was originally developed using the COVID-19 dataset, is reloaded and retrained on a newly merged dataset. The retraining process employs the same feature set utilized during the initial training of the COVID-19 model. By integrating and aligning features from distinct data sources, this approach enables the model to leverage the knowledge acquired from the COVID-19 dataset and apply it to a new dataset that includes monkeypox-related features. This method facilitates knowledge transfer and adaptation, allowing the model to generalize effectively across the expanded feature space derived from both the COVID-19 and monkeypox datasets.

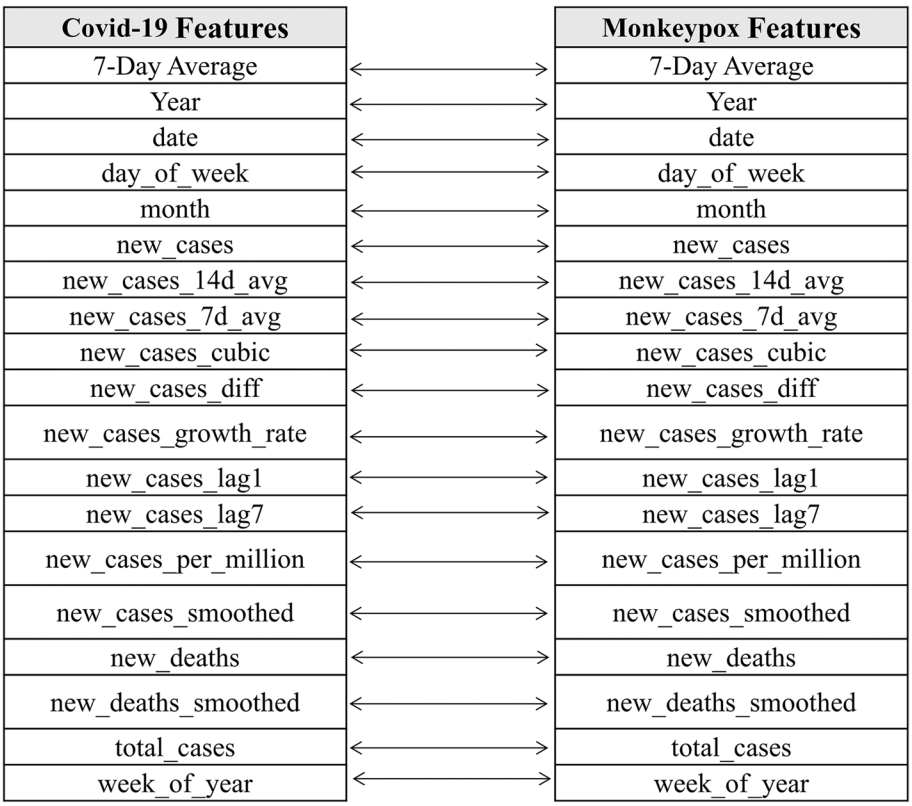


Fig. 2 Alignment of COVID-19 and Monkeypox dataset features

Incremental learning

Incremental learning is a technique that allows models to be gradually updated as data continue to change and increase [61]. It enables models to incorporate new data without retraining the entire model, thus maintaining real-time adaptability. The ridge regression, decision tree, XGBoost, and linear regression models used in this study do not support true incremental learning. To enable models to dynamically adapt to continuously changing new data in real time, we have devised a set of dynamic update mechanisms to approximate the effects of incremental learning.

Dynamic updating mechanism with sliding time windows

The sliding time window is a commonly used dynamic updating mechanism that involves sliding data within a fixed time window to progressively update a model [15]. Specifically, we define a fixed window of 30 days in length, where the model is trained using only the data within the window for updating. As new data arrives, the window slides forward to encompass the latest data while discarding the earliest data. The absolute error time series distribution plot of the blending model on the COVID-19 dataset (Fig. 3) clearly shows that the blending model can provide relatively accurate results for the next 7 days. Therefore, we slide the time window forward by 7 days each time to achieve a matching performance effect on the blending model, as shown in Fig. 4.

This study hypothesized that by utilizing 14 days of monkeypox data in the very early stages of an outbreak, followed by the incorporation of 16 days of COVID-19 data to

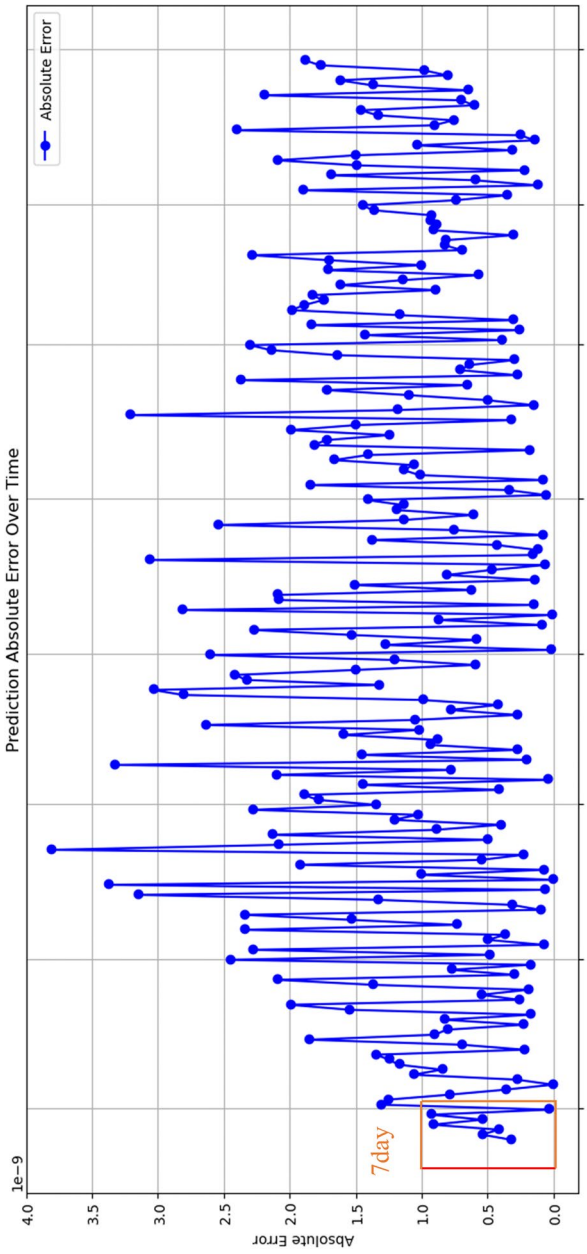


Fig. 3 presents the distribution of the absolute error time series predicted by the blending model. The x-axis spans from September 2022 to April 2023, whereas the y-axis represents the predicted absolute errors ranging from 0–4. The label "7 days" in a red box is positioned in the bottom left corner of the chart, highlighting a period of consistently low absolute errors

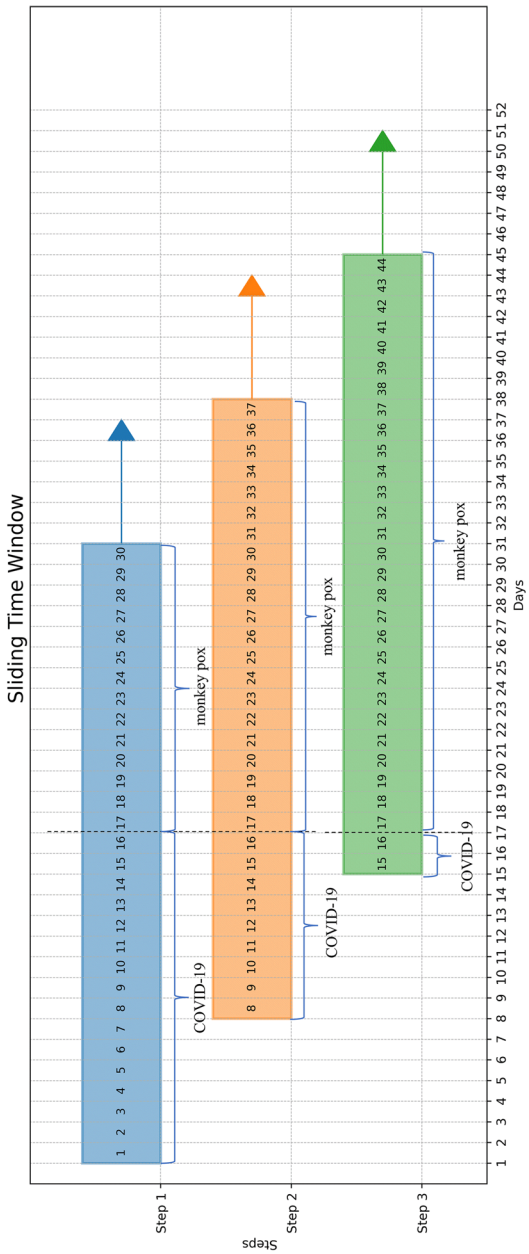


Fig. 4 Illustrates a sliding time window

complete a 30-day time window, with each additional week of monkeypox data replacing the earliest week of COVID-19 data, a sliding time window approach was implemented to enable real-time prediction and dynamic model updating [63].

The biological feature Rt is described below

In epidemiology, the effective reproduction number, Rt, is a crucial metric that signifies the average number of individuals to whom an infected person can transmit the infection during a specific period under prevailing conditions [8]. The dynamic fluctuations in Rt serve as indicators of the speed of transmission and the efficacy of control measures during an outbreak. Therefore, incorporating Rt as a biological characteristic is highly valuable for predicting the emergence of new infectious diseases such as monkeypox:

1. When Rt is greater than 1, the epidemic spreads, with each infected individual on average transmitting to more than one person.
2. When Rt equals 1, the epidemic is in a stable state, with the number of infections no longer increasing.
3. When Rt is less than 1, it signifies a decline in the epidemic, with each infected individual on average transmitting to fewer than one person.

Methodology for calculating Rt

The calculation of the effective reproduction number (Rt) is highly important in the formulation and assessment of public health policies such as social distancing measures and vaccination strategies, as it provides real-time insights into the impact of these interventions [46]. Various commonly employed methods for estimating Rt include the following:

1. Time series methods: This method uses time series data of reported case numbers to infer Rt by estimating the growth rate of infections. The growth rate is estimated by the slope of the logarithmically transformed case numbers, which is then combined with the virus's serial interval to calculate Rt.
2. The Bayesian method, a sophisticated statistical approach, integrates uncertainty and prior knowledge (such as historical data or other epidemiological characteristics). Typically, by employing Markov chain Monte Carlo (MCMC) techniques, this method estimates the probability distribution of Rt, offering confidence intervals and uncertainty assessments regarding Rt estimates.
3. Real time estimation tool:There are several readily available tools and software packages, such as EpiEstim and EpiFilter, that can be utilized for estimating Rt. These tools typically incorporate the aforementioned methods and enable researchers to input real-time case data to obtain estimates of Rt.

In this study, we utilized time series methods to calculate Rt to understand the transmission characteristics of monkepox and other related diseases. It is plausible to consider using a generation interval similar to that of smallpox or cowpox, which typically

falls between 12 and 14 days. For the purpose of this analysis, we may opt for an average value of 13 days as the generation interval to estimate Rt. The fundamental steps for calculating Rt are as follows:

The daily growth rate, denoted as r, is calculated by comparing the number of cases on two consecutive days. If Ct represents the number of cases on day t, the daily growth rate r can be calculated via the following formula:

1. The daily growth rate, denoted as r,is calculated; this estimation is derived by comparing the number of cases over two consecutive days. If we denote the number of cases on day t as Ct, then the daily growth rate r can be calculated via the following formula:

$$r_t = \ln\left(\frac{C_t}{C_{t-1}}\right)$$

   Here, 'ln' denotes the natural logarithm.
2. To calculate Rt, utilize the serial interval T in the context of secondary transmission;

Once we have the daily growth rate, we can calculate Rt via the estimated generation interval. The generation interval is the average time it takes for an individual to infect the next individual. The formula for Rt is as follows:

$$R_t = e^{r_t \times T}$$

Here, *e* represents the base of the natural logarithm, indicating that, in the absence of interventions, an infected individual will on average infect *Rt* other individuals during their infectious period.

Model evaluation metrics

When evaluating the performance of new infectious disease prediction models, selecting appropriate evaluation metrics is crucial. Accurate predictions can provide strong support for public health decision-making. By employing sliding time window technology, we can achieve real-time forecasting and dynamically update the model through blending [25]. Therefore, it is necessary to utilize multiple metrics to assess the predictive accuracy and reliability of the model comprehensively.

The root mean square error (RMSE) is the square root of the MSE, and serves as a metric for assessing the average magnitude of prediction errors. Being in the same units as the data facilitates the interpretation of the practical significance of errors. The calculation formula is as follows:

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

In this context, yi represents the actual value, $\widehat{y}$ i denotes the predicted value, and n represents the sample size. A lower RMSE value generally indicates higher model accuracy.

Table 3 Preliminary predictions of blending

Times	Number of training set	Number of validation set	Number of test set	RMSE	MAE
30 day	18 day	6 day	6 day	5.94	4.19
37 day	23 day	7 day	7 day	10.06	9.22
44 day	26 day	9 day	9 day	18.64	15.88
51 day	31 day	10 day	10 day	57.18	49.25

The mean absolute error (MAE) represents the average of the absolute errors between the predicted and actual values and measures the absolute magnitude of the errors regardless of their direction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right|$$

In this context, n represents the total number of observation points. yi represents the actual value observed at the ith point, with $\widehat{y}$ i indicating the corresponding predicted value. The absolute difference between the actual and predicted values is expressed as$\left|\mathrm{yi} - \widehat{\mathrm{y}}\mathrm{i}\right|$.

Results

In this section, we present the experimental results when the blending framework and transfer learning incremental learning technique are used on the monkeypox dataset. Our focus lies in evaluating the predictive performance of the models under different features and methodologies, encompassing the efficacy of time window-based feature transfer incremental learning and the performance enhancement upon the introduction of the biological feature Rt.

Preliminary predictions of the blending framework

Initially, we evaluated the progressive predictive capability of the blending framework on the monkeypox dataset. With the ongoing accumulation of early monkeypox data, there was a corresponding increase in the volume of data within the training set, enabling the observation of fluctuations in model performance. This methodology facilitated our comprehension of how the blending model performs under the circumstance of continually refreshing early data.

As shown in Table 3, the data were trained, validated, and tested within four time periods of 30, 37, 44, and 51 days as they increased. The corresponding numbers of days for the training, validation, and testing sets, along with their predictive performance metrics (RMSE and MAE), are presented below.

These findings clearly indicate that as the size of the training dataset increases, the model's prediction error significantly increases. This could be attributed to several factors:

1. The complexity of features in the training and testing data of the model: The blending model demonstrates promising performance in training on COVID-19 data and

Table 4 Compares the effects of blending transfer learning, incremental learning, and the prediction of monkeypox with the introduction of the biological feature Rt

Model Type	Times	RMSE	MAE
Transfer Learning Incremental Learning	Slide 0 weeks	3.14	2.3
Transfer Learning Incremental Learning Rt	Slide 0 weeks	3.14	2.3
Transfer Learning Incremental Learning	Slide 1 weeks	2.49	1.94
Transfer Learning Incremental Learning Rt	Slide 1 weeks	0.82	0.67
Transfer Learning Incremental Learning	Slide 2 weeks	0.93	0.61
Transfer Learning Incremental Learning Rt	Slide 2 weeks	0.45	0.31
Transfer Learning Incremental Learning	Slide 3 weeks	0.27	0.25
Transfer Learning Incremental Learning Rt	Slide 3 weeks	0.21	0.19

making predictions on COVID-19 datasets. However, as the volume of monkeypox data increases, the model may struggle to capture all features and patterns during training. Consequently, the model may exhibit signs of underfitting, leading to an increase in errors on the validation and test sets.

2. Variations in data distribution: The blending model is trained on a COVID-19 dataset, and over time, the monkepox epidemic may exhibit varying patterns of transmission. Discrepancies in the data distribution between the training and validation sets may hinder the model's ability to accurately predict future trends.

Comparing the effects of blending transfer learning with incremental learning for predicting monkeypox with the incorporation of the biological feature Rt

In this study, we further evaluated a blending model that demonstrated robust performance in training on COVID-19 data, utilizing transfer learning and incremental learning on the monkeypox dataset. To increase the predictive accuracy of the model, we introduced the biological feature Rt to capture the dynamic changes in epidemic spread. The model's prediction results with and without the Rt feature are presented in Table 4 for varying sliding time windows.

By employing transfer learning, we utilize the pertinent features of the COVID-19 dataset to construct a model for the monkeypox data. Incremental learning facilitates the model's progressive updates to accommodate the evolving data. The findings reveal a gradual reduction in prediction errors as the time window shifts, underscoring the adaptiveness of incremental learning to the ongoing integration of new data.

Introducing the effect of the biological trait Rt

Among all the tested models, the inclusion of the biological feature Rt demonstrated a significant performance improvement in each sliding window phase. In particular, following the first week of sliding, the model's RMSE decreased significantly from 2.49–0.82, and the MAE decreased from 1.94–0.67. These results indicate that the Rt feature substantially enhances the model's ability to capture epidemic transmission trends. Transfer learning leverages features from previous COVID-19 datasets, resulting in notable performance gains in data-scarce scenarios. The initial findings show that

at sliding week 0, the model's RMSE and MAE were 3.14 and 2.30, respectively. As the sliding window progresses, incremental learning gradually updates the model to adapt to new monkepox data. We observed a significant decrease in the model's RMSE and MAE with each additional week of data, highlighting the effectiveness of incremental learning in adapting to data changes. The introduction of Rt led to improved predictive performance across all time windows, particularly in the sliding windows of the first and second weeks. Rt, as a key biological feature, can reflect real-time changes in epidemic transmission; thus, its incorporation into the model significantly enhances prediction accuracy.

Discussion

In this study, we investigate the application and effectiveness of the blending framework, transfer learning, incremental learning, and biological feature Rt in predicting emerging infectious monkeypox. The experimental results demonstrate the significant advantages of these methods in addressing the challenges of early-stage emerging infectious disease data scarcity and real-time updates.

The applicability of the blending framework in predicting emerging infectious diseases

The blending framework enhances the overall predictive performance by combining predictions from multiple base models. Specifically, when predicting the next 6 days, the model achieves low RMSE and MAE values of 5.94 and 4.19, respectively. However, with increasing data, particularly at 51 days, the forecasts for the next 10 days show significant increases in the RMSE and MAE to 57.18 and 49.25, respectively. This suggests that the blending model performs well in short-term forecasts (within 7 days). The findings indicate that the blending framework is effective for short-term predictions but may require further optimization for handling long time series data. Adjusting model complexity or incorporating additional data preprocessing steps, such as feature selection or dimensionality reduction, may help improve the predictive accuracy of long time series data [18, 33].

Advantages of transfer learning and incremental learning in the context of data scarcity and dynamic updates

The application of transfer learning and incremental learning in this study demonstrates their potential in addressing data scarcity and real-time dynamic updates [49, 50]. The experimental results show that by transferring features from COVID-19 data to the monkeypox dataset, the model can still provide relatively accurate predictions in the early stages of data scarcity. For instance, at week 0 of the sliding window, the model's RMSE and MAE are 3.14 and 2.30, respectively, highlighting the advantage of transfer learning in leveraging existing knowledge for new tasks. Furthermore, incremental learning techniques enable the model to adapt to new data in real time, maintaining prediction accuracy. With each week of increase in the data, the model's prediction errors significantly decrease, reaching RMSE and MAE values of 0.27 and 0.25, respectively, after 3 weeks of sliding. This indicates the effectiveness of incremental learning in

real-time data updates, particularly in dealing with continuously changing data distributions. These results underscore the key advantages of transfer learning and incremental learning: they can initiate predictions in data-scarce scenarios and continuously update and optimize the model as new data arrives. This is particularly crucial for predicting new infectious diseases, as early-stage data are often limited and dynamically changing.

The advantages of the biological feature Rt in the early prediction of emerging infectious diseases

The inclusion of the biological feature Rt significantly enhances the predictive performance of the model, particularly in the very early stages of an epidemic. Rt, a key epidemiological indicator, reflects the average number of individuals to whom an infected person can transmit the virus under existing conditions [3, 39, 40]. The experimental data demonstrate that incorporating Rt results in a significant decrease in both the RMSE and the MAE of the model,for example, during a one-week sliding window, the RMSE decreases from 2.49 to 0.82, and the MAE decreases from 1.94 to 0.67. This indicates the effectiveness of the Rt feature in capturing the dynamics of epidemic spread, providing the model with more precise trend information. In the initial stages of a novel infectious disease outbreak, data are often limited and unstable; therefore, Rt can serve as a valuable feature, aiding the model in rapidly adapting to new circumstances and offering more accurate predictions.

Feasibility of cross-disease prediction

Comparison with recent research [28, 29, 59, 79, 81, 87]. This study combines integrated learning, incremental learning, and transfer learning, while introducing the biological feature Rt. This approach enables the model to be robust, real-time, and cross-disease predictive, even under conditions of data scarcity during the early stages of emerging infectious diseases. Such capabilities are critical in the initial phases of rapidly evolving epidemics. Despite the significant differences in transmission modes between COVID-19 and monkeypox, predictions for COVID-19 can be effectively adapted for monkeypox through feature generation, extraction, and alignment. The incremental learning technique allows for continuous adjustment of the model as outbreak data accumulate, enabling it to adapt to the specific transmission dynamics of monkeypox. This approach not only addresses the issue of data scarcity but also enhances the robustness and adaptability of the model in practical applications. The reproduction number (Rt), a common indicator of disease transmission capacity, can be effectively integrated into various infectious disease prediction models. Its incorporation improves the model's ability to capture the epidemic's spread rate, thereby enhancing prediction accuracy.

Potential applications of models in public health decision-making

The blending prediction model developed in this study holds significant potential for application in public health decision-making. By integrating transfer learning, incremental learning, and the biological feature Rt, we can provide highly accurate early epidemic forecasts for emerging infectious diseases, which is crucial for resource allocation and emergency response. For example, the model can predict the growth trend of case numbers in the next 7 days, aiding health authorities in preparing medical resources in

advance, formulating isolation policies, and optimizing vaccine distribution strategies. Accurate predictions can significantly reduce public health risks, enhancing the timeliness and effectiveness of prevention and control measures. Compared with previous studies, this research has made several important advancements. First, we introduce the blending framework, which synthesizes the strengths of various models when dealing with multiple model outputs, thereby enhancing the overall predictive performance. Second, through transfer learning and incremental learning, we successfully leveraged knowledge from COVID-19 data, significantly improving the accuracy of monkepox epidemic prediction. Finally, the incorporation of the biological feature Rt provides profound insights into the dynamics of epidemic spread, enabling the model to more accurately capture changes in disease transmission trends. These enhancements position our model as superior to many traditional methods in terms of predictive accuracy and adaptability.

Insights for the development of future infectious disease prediction tools

The results of this study provide several important insights for the development of future infectious disease prediction tools. First, the integration of multiple models (such as the blending framework) is an effective method for improving prediction accuracy, particularly when dealing with complex and variable data [75]. Second, transfer learning performs well in cases of data scarcity, indicating that leveraging existing relevant data for knowledge transfer can significantly enhance model performance [79, 81]. Incremental learning enables real-time prediction and dynamic model updates. Third, the incorporation of biological features (such as Rt) can provide crucial epidemiological information to the model, aiding in capturing potential changes in outbreaks. Future research could further explore the optimization of these methods. For example, more advanced feature selection and extraction techniques should be investigated to further enhance the predictive capabilities of the models. Additionally, the development of real-time data updating and automated model tuning systems will make prediction tools more efficient and reliable in practical applications.

Limitations of the research

One significant limitation of this study lies in the impact of data quality and quantity on the model's performance. The data utilized originate primarily from public sources such as the monkepox dataset provided by Our World in Data. These datasets may suffer from inconsistencies in data collection, reporting delays, or missing information, all of which can potentially affect the predictive accuracy of the model [13, 22, 52]. For example, in cases of sparse data, the model may struggle to adequately capture the characteristics of disease transmission, leading to increased prediction errors. Furthermore, the temporal span and geographical coverage of the data also influence the model's generalizability. During the early stages of an epidemic, data are typically limited and unstable, which could result in model overfitting to the restricted training data and failure to accurately forecast future trends. Therefore, the quality and quantity of data are critical factors influencing model performance, particularly in addressing rapidly evolving infectious disease outbreaks [17, 65].

Another limitation is the generalizability of the model in predicting different diseases. While this study improved the prediction accuracy of monkeypox outbreaks by utilizing features from COVID-19 data through transfer learning, the generalizability of this approach may be limited. Different diseases have distinct transmission mechanisms and characteristics, such as routes of transmission, incubation periods, and severities of infection. Therefore, the successful application of a model for one disease does not guarantee similar performance for other diseases. Further research is needed to explore ways to enhance the model's generalization ability to better adapt to various infectious disease scenarios. This may involve developing more universal feature extraction methods or incorporating more biological and epidemiological knowledge into the model. Additionally, integrating multiple data sources, such as epidemiological survey data, genomic data, and environmental data, may help improve the model's prediction accuracy and generalizability [83].

Conclusion

This study investigates the application of the blending framework trained with COVID -19 data, combined with transfer learning and incremental learning, in the prediction of monkepox outbreaks. The incorporation of the biological feature Rt significantly enhances the predictive accuracy of the model. The research findings indicate the following:

1. Evaluation of the blending framework's performance and limitations: The blending framework proves effective in short-term forecasting but may require further optimization when handling long time series data. Adjusting model complexity or incorporating additional data preprocessing steps, such as feature selection or dimensionality reduction, could enhance the predictive accuracy of long time series data.
2. The advantages of transfer learning: By leveraging the features of a COVID-19 dataset, we successfully applied it to predict monkepox outbreaks. The utilization of transfer learning significantly enhanced the model's predictive capacity in scenarios of limited data availability, underscoring the critical importance of leveraging interdisciplinary knowledge in forecasting emerging infectious diseases.
3. Dynamic adaptability of incremental learning: Through incremental learning, models can dynamically adapt to changes in new data, maintaining a high level of predictive accuracy. This technique is particularly suitable for situations involving dynamic changes, such as in the case of epidemic outbreaks, enabling real-time forecast updates.
4. Introduction of the biological feature Rt: The inclusion of the effective reproduction number (Rt) as a crucial indicator reflecting the dynamics of disease spread significantly enhances the predictive ability of models. Particularly in the early stages of emerging infectious diseases, incorporating Rt assists models in capturing changes in disease transmission more accurately.

This study demonstrates the potential application of a blending framework that integrates transfer learning, incremental learning, and biological feature Rt in infectious

disease prediction, offering reliable technological support for public health emergency responses.

Research in the future

While this study yielded valuable results, there are still numerous areas that warrant further exploration. The following are some research recommendations for future investigations in the field of forecasting emerging infectious diseases:

1. Improving Data Quality and Integrating Multiple Data Sources: Future research should focus on enhancing the quality and usability of data, particularly in situations where early epidemic data are scarce. Integrating multiple data sources, such as epidemiological, environmental, and genomic data, will aid in the construction of more comprehensive and precise predictive models.
2. Real-time prediction and automated updating system: With the continuous influx of new data, the development of real-time prediction and automated model updating systems is becoming increasingly crucial. Such systems should be capable of automatically acquiring new data, updating model parameters, and generating prediction outcomes to provide timely information support to decision-makers.
3. Interdisciplinary collaboration: The prediction of emerging infectious diseases involves knowledge from various fields, such as epidemiology, statistics, and data science. Future research should emphasize interdisciplinary collaboration, integrating the latest findings from different disciplines to develop more precise and practical forecasting models.

Through further research and exploration, we can continuously enhance the accuracy and practicality of infectious disease prediction models, thus providing more robust support for global public health.

Acknowledgements
This part is not applicable.

Authors' contributions
D.W. and H.H.K wrote the main text of the manuscript and D. drew Figs. 1- 2. All authors reviewed the manuscript.

Funding
This part is not applicable.

Data availability
The U.S. monkeypox dataset that supports the conclusions of this article is available at https://ourworldindata.org/mpox.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors approved the manuscript and agreed with its publication.

Competing interests
The authors declare no competing interests.

Received: 31 July 2024 Accepted: 9 October 2024
Published online: 22 October 2024

References

1. Alizargar A, Chang YL, Alkhaleefah M, Tan TH. Precision Non-Alcoholic Fatty Liver Disease (NAFLD) Diagnosis: Leveraging Ensemble Machine Learning and Gender Insights for Cost-Effective Detection. Bioengineering (Basel). 2024;11(6):600.
2. Alqaissi EY, Alotaibi FS, Ramzan MS. Modern machine-learning predictive models for diagnosing infectious diseases. Comput Math Methods Med. 2022;2022(1):1–13. https://doi.org/10.1155/2022/6902321.
3. Alvarez L, Colom M, Morel JD, Morel JM. Computing the daily reproduction number of COVID-19 by inverting the renewal equation using a variational technique. Proc Natl Acad Sci. 2021;118(50):e2105112118.
4. Anjum M, Saher R, Saeed MN. Optimizing type 2 diabetes management: AI-enhanced time series analysis of continuous glucose monitoring data for personalized dietary intervention. PeerJ Computer science. 2024;10: e1971.
5. Błażkiewicz M. Evaluation of geometric attractor structure and recurrence analysis in professional dancers. Entropy (Basel). 2022;24(9):1310.
6. Borré A, Seman LO, Camponogara E, Stefenon SF, Mariani VC, Coelho LDS. Machine fault detection using a hybrid CNN-LSTM attention-based model. Sensors. 2023;23(9):4512.
7. Canino MP, Cesario E, Vinci A, Zarin S. Epidemic forecasting based on mobility patterns: an approach and experimental evaluation on COVID-19 Data. Soc Netw Anal Min. 2022;12(1):116.
8. Champredon D, Papst I, Yusuf W. ern: An [Formula: see text] package to estimate the effective reproduction number using clinical and wastewater surveillance data. PLoS ONE. 2024;19(6): e0305550.
9. Chang H, Esteves IM, Neumann AR, Mohajerani MH, McNaughton BL. Cortical reactivation of spatial and non-spatial features coordinates with hippocampus to form a memory dialogue. Nat Commun. 2023;14(1):7748.
10. Chaudhary V, Khanna V, Ahmed Awan HT, Singh K, Khalid M, Mishra YK, Kaushik A. Towards hospital-on-chip supported by 2D MXenes-based 5th generation intelligent biosensors. Biosensors and bioelectronics. 2023;220: 114847.
11. Chen J, Guo C, Lu M, Ding S. Unifying diagnosis identification and prediction method embedding the disease ontology structure from electronic medical records. Front Public Health. 2021;9: 793801.
12. Choi SH, Park JK, An D, Kim CH, Park G, Lee I, Lee S. Fault Diagnosis Method for Human Coexistence Robots Based on Convolutional Neural Networks Using Time-Series Data Generation and Image Encoding. Sensors. 2023;23(24):9753.
13. Ciobanu-Caraus O, Aicher A, Kernbach JM, Regli L, Serra C, Staartjes VE. A critical moment in machine learning in medicine: on reproducible and interpretable learning. Acta Neurochir. 2024;166(1):14.
14. Cui L, Agrawal A. Special supplement issue on quality assurance and enrichment of biological and biomedical ontologies and terminologies. BMC Med Inform Decis Mak. 2024;23(Suppl 1):302.
15. Cui R, Hua W, Qu K, Yang H, Tong Y, Li Q, Liu C. An interpretable early dynamic sequential predictor for sepsis-induced coagulopathy progression in the real-world using machine learning. Frontiers in medicine. 2021;8: 775047.
16. Cunningham GB, Watanabe NM, Buzuvis E. Anti-transgender rights legislation and internet searches pertaining to depression and suicide. PLoS ONE. 2022;17(12): e0279420.
17. De Salazar PM, Lu F, Hay JA, Gómez-Barroso D, Fernández-Navarro P, Martínez EV, Hernán MA. Near real-time surveillance of the SARS-CoV-2 epidemic with incomplete data. PLoS computational biology. 2022;18(3): e1009964.
18. Deng Z, Zhang J, Li J, Zhang X. Application of deep learning in plant-microbiota association analysis. Front Genet. 2021;12: 697090.
19. Didier AJ, Nigro A, Noori Z, Omballi MA, Pappada SM, Hamouda DM. Application of machine learning for lung cancer survival prognostication-A systematic review and meta-analysis. Frontiers in artificial intelligence. 2024;7: 1365777.
20. Ding D, Zhang R. China's COVID-19 control strategy and its impact on the global pandemic. Front Public Health. 2022;10:857003.
21. Du H, Yu M, Xue H, Lu X, Chang Y, Li Z. Association between sarcopenia and cognitive function in older Chinese adults: Evidence from the China health and retirement longitudinal study. Front Public Health. 2022;10:1078304.
22. Du M, Huang X, Li S, Xu L, Yan B, Zhang Y, Liu X. A nomogram model to predict malignant cerebral edema in ischemic stroke patients treated with endovascular thrombectomy: an observational study. Neuropsychiatric disease and treatment. 2020;16:2913–20.
23. El Taha L, Beyrouthy C, Tamim H, Ghazeeri G. Knowledge and attitudes among Lebanese pregnant women and women seeking fertility treatment during the COVID-19 outbreak: a cross-sectional survey. BMJ Open. 2022;12(3): e057873.
24. Estiri H, Strasser ZH, Klann JG, Naseri P, Wagholikar KB, Murphy SNJN, d. m. Predicting COVID-19 mortality with electronic medical records. NPJ digital medicine. 2021. https://doi.org/10.1038/s41746-021-00383-x.
25. Goetz C, Humm B. Decentralized real-time anomaly detection in cyber-physical production systems under industry constraints. Sensors (Basel). 2023;23(9):4207.
26. Gong H, Wang M, Zhang H, Elahe MF, Jin M. An explainable AI approach for the rapid diagnosis of COVID-19 Using Ensemble Learning Algorithms. Front Public Health. 2022;10: 874455.
27. Guo X, Yang W, Xiong X, Wang Z, Zou X. MEMS reservoir computing system with stiffness modulation for multi-scene data processing at the edge. Microsyst Nanoeng. 2024;10:84.
28. Huang L, Li OZ, Yin X. Inferring China's excess mortality during the COVID-19 pandemic using online mourning and funeral search volume. Sci Rep. 2023;13(1):15665.
29. Huang Y, Zhang P, Wang Z, Lu Z, Wang ZJNPL. HFMD cases prediction using transfer one-step-ahead learning. 2023;55(3):2321–39.
30. Ji B, Pi W, Liu W, Liu Y, Cui Y, Zhang X, Peng S. HyperVR: a hybrid deep ensemble learning approach for simultaneously predicting virulence factors and antibiotic resistance genes. NAR genomics and bioinformatics. 2023;5(1): lqad012.
31. Jiang Y, Li Q, Trevisan G, Linhares DCL, MacKenzie C. Investigating the relationship of porcine reproductive and respiratory syndrome virus RNA detection between adult/sow farm and wean-to-market age categories. PLoS ONE. 2021;16(7): e0253429.
32. Jorge DCP, Oliveira JF, Miranda JGV, Andrade RFS, Pinho STR. Estimating the effective reproduction number for heterogeneous models using incidence data. Royal Society open science. 2022;9(9): 220005.

33. Kang Z, Fan R, Zhan C, Wu Y, Lin Y, Li K, et al. The Rapid non-destructive differentiation of different varieties of rice by fluorescence hyperspectral technology combined with machine learning. Molecules (Basel). 2024;29(3):682.
34. Keshavamurthy R, Charles LE. Predicting Kyasanur forest disease in resource-limited settings using event-based surveillance and transfer learning. Sci Rep. 2023;13(1):11067.
35. Kwon B, Son H. Accurate Path Loss Prediction Using a Neural Network Ensemble Method. Sensors. 2024;24(1):304.
36. Larsen SL, Kraay ANM. Transparent transmission models for informing public health policy: the role of trust and generalizability. Proceedings Biological sciences. 2024;291(2015):20232273.
37. Lee J, Kim JN, Dallan LAP, Zimin VN, Hoori A, Hassani NS, Wilson DL. Deep learning segmentation of fibrous cap in intravascular optical coherence tomography images. Scientific reports. 2024;14(1):4393.
38. Lenatti M, Narteni S, Paglialonga A, Rampa V, Mongelli M. Dual-view single-shot multibox detector at urban intersections: Settings and performance evaluation. Sensors. 2023;23(6):3195.
39. Li H, Yang Y, Chen J, Li Q, Chen Y, Zhang Y, Xiang J. Epidemiological characteristics of overseas-imported infectious diseases identified through airport health-screening measures: a case study on Fuzhou, China. Trop Med Infect Dis. 2024;9(6):138.
40. Li X, Patel V, Duan L, Mikuliak J, Basran J, Osgood ND. Real-Time Epidemiology and Acute Care Need Monitoring and Forecasting for COVID-19 via Bayesian Sequential Monte Carlo-Leveraged Transmission Models. Int J Environ Res Public Health. 2024;21(2):193. https://doi.org/10.3390/ijerph21020193.
41. Liu C, Su H. Prediction of martensite start temperature of steel combined with expert experience and machine learning. Sci Technol Adv Mater. 2024;25(1):2354655.
42. Liu M, Liu Y, Liu J. Epidemiology-aware Deep Learning for Infectious Disease Dynamics Prediction. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, vol. 33. New York: Association for Computing Machinery; 2023. p. 4084–8. https://doi.org/10.1145/3583780.3615139.
43. Liu T, Huang J, He Z, Zhang Y, Yan N, Zhang CJP, Ming W-K. Rethinking of value of early-stage infectious disease modelling to public health: a real-world data validation of SIR models. 2022. https://doi.org/10.21203/rs.3.rs-2069781/v1.
44. Liu T, Huang J, He Z, Zhang Y, Yan N, Zhang CJP, Ming W-K. A real-world data validation of the value of early-stage SIR modelling to public health. Sci Rep. 2023;13(1):9164. https://doi.org/10.1038/s41598-023-36386-9.
45. Lou J, Wang B, Li J, Ni P, Jin Y, Chen S, Duan G. The CRISPR-Cas system as a tool for diagnosing and treating infectious diseases. Molecular biology reports. 2022;49(12):11301–11.
46. Madewell ZJ, Yang Y, Longini IM, Halloran ME, Vespignani A, Dean NE. Rapid review and meta-analysis of serial intervals for SARS-CoV-2 Delta and Omicron variants. BMC Infect Dis. 2023;23(1):429.
47. Mei H, Peng J, Wang T, Zhou T, Zhao H, Zhang T, Yang Z. Overcoming the limits of cross-sensitivity: pattern recognition methods for chemiresistive gas sensor array. Nano-micro letters. 2024;16(1):269.
48. Nair A, Ahirwar A, Singh S, Lodhi R, Lodhi A, Rai A, Vinayak V. Astaxanthin as a king of ketocarotenoids: structure, synthesis, accumulation, bioavailability and antioxidant properties. Marine Drugs. 2023;21(3):176.
49. Narkhede P, Walambe R, Poddar S, Kotecha K. Incremental learning of LSTM framework for sensor fusion in attitude estimation. PeerJ Computer science. 2021;7: e662.
50. Noordman CR, Yakar D, Bosma J, Simonis FFJ, Huisman H. Complexities of deep learning-based undersampled MR image reconstruction. European radiology experimental. 2023;7(1):58.
51. Olum R, Ahaisibwe B, Atuhairwe I, Balizzakiwa T, Kizito P, Apiyo M, Kalanzi J, Nabawanuka A, Bahatungire R, Kerry V. Readiness To Manage Ebola Virus Disease Among Emergency Healthcare Workers in Uganda: A Nationwide Multicenter Survey. 2024. https://doi.org/10.21203/rs.3.rs-4212996/v1.
52. Ortiz-Barrios M, Petrillo A, Arias-Fonseca S, McClean S, de Felice F, Nugent C, Uribe-López SA. An AI-based multiphase framework for improving the mechanical ventilation availability in emergency departments during respiratory disease seasons: a case study. Int J Emerg Med. 2024;17(1):45.
53. Perichart-Perera O, Avila-Sosa V, Solis-Paredes JM, Montoya-Estrada A, Reyes-Muñoz E, Rodríguez-Cano AM, González-Leyva CP, Sánchez-Martínez M, Estrada-Gutierrez G, Irles C. Vitamin D deficiency, excessive gestational weight gain, and oxidative stress predict small for gestational age newborns using an artificial neural network model. Antioxidants (Basel). 2022;11(3):574.
54. Perramon-Malavez A, Bravo M, de Rioja VL, Català M, Alonso S, Álvarez-Lacalle E, Prats C. A semi-empirical risk panel to monitor epidemics: multi-faceted tool to assist healthcare and public health professionals. Frontiers in public health. 2023;11:1307425.
55. Popescu S, Myers N. Interdisciplinary information for infectious disease response: exercising for improved medical/public health communication and collaboration. Disaster Med Public Health Prep. 2021;15(5):546–50.
56. Qiu H, Luo L, Su Z, Zhou L, Wang L, Chen Y. Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure. BMC Med Inform Decis Mak. 2020;20(1):83.
57. Raza S, Schwartz B. Constructing a disease database and using natural language processing to capture and standardize free text clinical information. Sci Rep. 2023;13(1):8591.
58. Rodrigues J, Studer E, Streuber S, Meyer N, Sandi C. Locomotion in virtual environments predicts cardiovascular responsiveness to subsequent stressful challenges. Nat Commun. 2020;11(1):5904.
59. Roster K, Connaughton C, Rodrigues FA. Forecasting new diseases in low-data settings using transfer learning. Chaos Solitons Fractals. 2022;161:112306.
60. Ruan Y, Huang T, Zhou W, Zhu J, Liang Q, Zhong L, Xie Y. The lead time and geographical variations of Baidu Search Index in the early warning of COVID-19. Scientific reports. 2023;13(1):14705.
61. Shyaa MA, Zainol Z, Abdullah R, Anbar M, Alzubaidi L, Santamaría J. Enhanced Intrusion Detection with Data Stream Classification and Concept Drift Guided by the Incremental Learning Genetic Programming Combiner. Sensors (Basel). 2023;23(7):3736.
62. Soliman M, Lyubchich V, Gel YR. Complementing the power of deep learning with statistical model fusion: Probabilistic forecasting of influenza in Dallas County, Texas, USA. Epidemics. 2019;28: 100345.

63. Soni M, Khan IR, Basir S, Chadha R, Alguno AC, Bhowmik T. Light-weighted deep learning model to detect fault in IoT-based industrial equipment. Comput Intell Neurosci. 2022;2022:2455259.
64. Su K, Yuan X, Huang Y, Yuan Q, Yang M, Sun J, Yuan Z. Improved prediction of knee osteoarthritis by the machine learning model XGBoost. Indian journal of orthopaedics. 2023;57(10):1667–77.
65. Vobugari N, Raja V, Sethi U, Gandhi K, Raja K, Surani SR. Advancements in oncology with artificial intelligence—a review article. Cancers. 2022;14(5):1349.
66. Wang H, Cui W, Guo Y, Du Y, Zhou Y. Machine Learning Prediction of Foodborne Disease Pathogens: Algorithm Development and Validation Study. JMIR Med Inform. 2021;9(1): e24924.
67. Wang H, Qiu J, Li C, Wan H, Yang C, Zhang T. Applying the spatial transmission network to the forecast of infectious diseases across multiple regions. Front Public Health. 2022;10: 774984.
68. Wang J, Zhang H, Chen N, Zeng T, Ai X, Wu K. PorcineAI-enhancer: prediction of pig enhancer sequences using convolutional neural networks. Animals. 2023;13(18):2935.
69. Wang L, Liu Y, Chen H, Qiu S, Liu Y, Yang M, Du X, Li Z, Hao R, Tian H, Song H. Search-engine-based surveillance using artificial intelligence for early detection of coronavirus disease outbreak. J Big Data. 2023;10(1):169. https://doi.org/10.1186/s40537-023-00847-9.
70. Wang P, Zhang W, Wang H, Shi C, Li Z, Wang D, Hao Y. Predicting the incidence of infectious diarrhea with symptom surveillance data using a stacking-based ensembled model. BMC infectious diseases. 2024;24(1):265.
71. Wang Y, Cao Z, Zeng D, Wang X, Wang Q. Using deep learning to predict the hand-foot-and-mouth disease of enterovirus A71 subtype in Beijing from 2011 to 2018. Sci Rep. 2020;10(1):12201.
72. Wang Y, Xu C, Li Y, Wu W, Gui L, Ren J, Yao S. An advanced data-driven hybrid model of sarima-nnnar for tuberculosis incidence time series forecasting in Qinghai Province, China. Infection and drug resistance. 2020;13:867–80.
73. Wang Z, Zhang P, Huang Y, Chao G, Xie X, Fu YJAI. Oriented transformer for infectious disease case prediction. 2023;53(24):30097–112.
74. White L, Basurra S, Alsewari AA, Saeed F, Addanki SM. Temporal meta-optimiser based sensitivity analysis (TMSA) for agent-based models and applications in children's services. Sci Rep. 2024;14(1):9105.
75. Xie M, Lin S, Dong K, Zhang S. Short-Term Prediction of Multi-Energy Loads Based on Copula Correlation Analysis and Model Fusions. Entropy (Basel). 2023;25(9):1343.
76. Xu D, Chan WH, Haron HJPCS. Enhancing infectious disease prediction model selection with multi-objective optimization: an empirical study. 2024;10: e2217.
77. Yang J, Zhou J, Luo T, Xie Y, Wei Y, Mai H, Huang J. Predicting pulmonary tuberculosis incidence in China using Baidu search index: an ARIMAX model approach. Environmental health and preventive medicine. 2023;28:68.
78. Yang L, Li G, Yang J, Zhang T, Du J, Liu T, Zhang X, Han X, Li W, Ma L, Feng L. Deep-learning model for influenza prediction from multisource heterogeneous data in a megacity. J Med Internet Res. 2023;25:e44238.
79. Yang L, Yang J, He Y, Zhang M, Han X, Hu X, Li W, Zhang T, Yang W. Enhancing infectious diseases early warning: A deep learning approach for influenza surveillance in China. Prev Med Rep. 2024;43:102761.
80. Yang R, Lin Z, Cai Y, Chen N, Zhou Y, Zhang J, Hong G. Assessing the risk of prenatal depressive symptoms in Chinese women: an integrated evaluation of serum metabolome, multivitamin supplement intake, and clinical blood indicators. Front Psych. 2023;14:1234461.
81. Yang S, Cui L, Wang L, Wang T, You J. Enhancing multimodal depression diagnosis through representation learning and knowledge transfer. Heliyon. 2024;10(4): e25959.
82. Yoo HY, Lee KC, Woo JE, Park SH, Lee S, Joo J, Park BJ. A genome-wide association study and machine-learning algorithm analysis on the prediction of facial phenotypes by genotypes in Korean women. Clinical, cosmetic and investigational dermatology. 2022;15:433–45.
83. Yu Y, Tan J, Yang Y, Zhang B, Yao X, Sang S, Deng S. The differential diagnostic value of radiomics signatures between single-nodule pulmonary metastases and second primary lung cancer in patients with colorectal cancer. Technol Cancer Res Treat. 2023;22: 15330338231175735.
84. Zemariam AB, Yimer A, Abebe GK, Wondie WT, Abate BB, Alamaw AW, Ngusie HS. Employing supervised machine learning algorithms for classification and prediction of anemia among youth girls in Ethiopia. Scientific reports. 2024;14(1):9080.
85. Zhang M, Yang W, Chen D, Fu C, Wei F. AM-MSFF: A Pest Recognition Network Based on Attention Mechanism and Multi-Scale Feature Fusion. Entropy (Basel). 2024;26(5):431.
86. Zhang P, Wang Z, Huang Y, Wang MJK-BS. Dual-grained directional representation for infectious disease case prediction. 2022;256: 109806.
87. Zhou W, Huang D, Liang Q, Huang T, Wang X, Pei H, Qin L. Early warning and predicting of COVID-19 using zero-inflated negative binomial regression model and negative binomial regression model. BMC Infect Dis. 2024;24(1):1006.
88. Zhu D, Bu Q, Zhu Z, Zhang Y, Wang Z. Advancing autonomy through lifelong learning: a survey of autonomous intelligent systems. Front Neurorobot. 2024;18: 1385778.

Publisher's Note