# Time series forecasting of Valley fever infection in Maricopa County, AZ using LSTM

Xueting Jin,[a,b] Fangwu Wei,[a,*] Srinivasa Srivatsav Kandala,[a] Tejas Umesh,[c] Kayleigh Steele,[a] John N. Galgiani,[d,e] and Manfred D. Laubichler[a]

[a]Decision Theater, Knowledge Enterprise, Arizona State University, Tempe, AZ, USA
[b]School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ, USA
[c]Barrow Neurological Institute, St Joseph's Hospital, Phoenix, AZ, USA
[d]Valley Fever Center for Excellence and the Departments of Medicine and Immunobiology, University of Arizona College of Medicine-Tucson, Tucson, AZ, USA
[e]BIO5 Institute, University of Arizona, Tucson, AZ, USA

The Lancet Regional Health - Americas 2025;43: 101010

Published Online xxx https://doi.org/10.1016/j.lana.2025.101010

## Summary

**Background** Coccidioidomycosis (CM), also known as Valley fever, is a respiratory infection. Recently, the number of confirmed cases of CM has been increasing. Precisely defining the influential factors and forecasting future infection can assist in public health messaging and treatment decisions.

**Methods** We utilized Long Short-Term Memory (LSTM) networks to forecast CM cases, based on the daily pneumonia cases in Maricopa County, Arizona from 2020 to 2022. Besides weather and climate variables, we examined the impact of people's lifestyle change during COVID-19. Factors, including temperature, precipitation, wind speed, $PM_{10}$ and $PM_{2.5}$ concentration, drought, and stringency index, were included in LSTM networks, considering their association with CM prevalence, time-lag effect, and correlation with other factors.

**Findings** LSTM can predict CM prevalence with accurate trend and low mean squared error (MSE). We also found a tradeoff between the length of the forecasting period and the performance of the forecasting model. The models with longer forecasting periods have less accurate trends over time and higher MSEs. Two models with different lengths of forecasting periods, 10 days and 30 days, are identified with good prediction.

**Interpretation** LSTM algorithms, combined with traditional statistical methods, could help with the forecasting of CM cases. By predicting the CM prevalence, our results can inform researchers, epidemiologists, clinicians, and the public in order to assist public health.

**Funding** "Getting to the Source of Arizona's Valley Fever Problem: A Tri-University Collaboration to Map and Characterize the Pathogen Where It Grows" funded by the Arizona Board of Regents.



**Keywords:** Coccidioidomycosis; Valley fever; Deep learning; LSTM; Time series forecasting

## Introduction

Coccidioidomycosis (CM), also known as Valley fever, is a respiratory infection caused by inhaling spores from the fungi *Coccidioides immitis* and *Coccidioides posadasii*.[1] These fungi are endemic to the southwestern United States, as well as parts of Mexico and Central and South America.[2,3] *Coccidioides* spp. grow in soil and can be inhaled when the spores become airborne. Although over half of infections do not cause illness, the remainder will experience a variety of symptoms, most commonly those of community acquired pneumonia (CAP), which last from weeks up to months. A small percentage will experience even more severe illness, which is potentially life-threatening.[4,5] These symptoms can be easily confused with other diseases and delays in diagnosis are common without specific testing.[6,7]

The Centers for Disease Control and Prevention (CDC) recorded 20,003 cases of CM in 2019, most of which occurred in California or Arizona. In 2022, in Arizona alone, over 9500 cases were reported with 75 resulting in death. These case reports also likely underestimate the actual incidence of this illness as many patients may not be tested due to the flu-like presentation of symptoms.[8] According to the CDC, CM is

*Corresponding author. Arizona State University, Tempe, AZ, USA.
*E-mail address:* fwei15@asu.edu (F. Wei).

## Research in context

### Evidence before this study

We searched online Google Scholar and Web of Science with no language restrictions using the terms "Valley fever", "Coccidioidomycosis", "influence factor", and "prediction". Most studies identified used data-driven approaches such as regression analysis to explore the relationship between the coccidioidomycosis (CM) incidence and the associated factors and forecast the future incidence trend when considering the factors from weather and climate, soil, landform, land cover and socio-demographic aspects. These studies struggled with nonlinear relationships, unpredictable variables like weather, and temporal variations, which highlight the need for more effective models that address these challenges to improve forecasting accuracy. By introducing LSTM, this study provides a new method of forecasting CM infection in time series with accurate trend prediction.

### Added value of this study

To the best of our knowledge, this study is the first to use a deep learning algorithm to forecast CM cases. To develop a better understanding of the potential influence variables and their impact on CM infection, we propose a methodology of combining traditional statistical analysis with a LSTM algorithm. Factors to be included in LSTM networks, including weather, air quality and stringency index, are selected using statistical analysis, consideration of their association with CM prevalence, time lag effect, and correlation with other factors. Leveraging deep learning capabilities, we introduce a general analytic framework as future work to facilitate the CM research.

### Implications of all the available evidence

Our research found that LSTM can predict CM prevalence with accurate trend and low mean squared error. The methodology we propose can assist in precisely defining the influential factors and forecasting future infection of CM. The CM community, including researchers, epidemiologists, clinicians, and the public, can benefit from this study to support public health in a planned and economical manner, especially during future shocks as seen during the COVID-19 pandemic response.

estimated to account for 15%–30% of community-acquired cases of CAP in highly endemic areas, such as Maricopa County in Arizona. The number of reported cases of CM has also steadily increased since 2016. For example, in California, Valley fever cases tripled from 2014 to 2018, and around 8000 cases were reported each year from 2018 to 2022.[9] Confirmed cases in Arizona almost doubled from 2017 to 2021.[10]

Understanding factors associated with CM infection and predicting the trend in the future can be useful for planning, decision-making, and prevention. The early prediction of epidemics will benefit government and healthcare departments by enabling a timely response to outbreaks. Additionally, it will minimize the negative impact and ensure the use of resources in a planned manner. Studies have been conducted to understand the factors associated with CM infection. The results draw connections between CM incidence and factors such as weather and climate,[2,4,11–18] soil,[2,5,11,13–16,18,19] landform,[2,15,16] land cover,[2,15,18] and socio-demographic variables.[3,16,20]

As for the analytic methods, we found most current studies use data-driven approaches such as autoregressive integrated moving average (ARIMA) models and other regression models to explore the relationship between the potential factors and the CM incidence and forecast the future incidence trend.[17,20] ARIMA models are prevalent in time series data modeling and forecasting, but they have some major limitations. For instance, the accuracy of prediction results largely depends on the prediction accuracy of the independent variables.[21] However, variables, such as weather, are becoming more unpredictable recently as the increasing frequency and intensity of extreme weather events are documented.[22] Also, it is assumed that there is a constant standard deviation in errors in ARIMA model,[23] which may not be satisfied in practice. Furthermore, in a simple ARIMA model, it is hard to model the nonlinear relationships between variables and forecast real-time transmission.[24] Such research usually and often neglects the temporal variation of the infection.[25]

Other studies consider the temporal variation of the CM infection,[3,18] for example by dividing the entire data into several sub-datasets based on the seasons/months and then building regression models for each sub-dataset.[4,11,15] The results reflect that the impact of weather and climate variables on CM infection varied over time. For example, precipitation during the hottest and driest parts of the year (April through June), as opposed to wetter seasons, was found to be most favorable for Coccidioides growth in the environment.[4] Some other studies used the generalized additive model (GAM)[16]; a type of regression model that allows for the creation of smooth, non-linear functions to model the relationship between the response variable and one or more predictor variables. However, GAMs have some limitations which may compromise the performance of the analysis/forecasting results.[26] One limitation is that GAMs may not be able to handle very large or high-dimensional datasets well, as they require more computational resources and may suffer from overfitting or multicollinearity. Another limitation is that GAMs may not capture all the features or patterns of the data, such as outliers, clusters, or heterogeneity. The third limitation is that GAMs may not account for

all uncertainty/variability in the dataset, such as measurement error, missing data, or causal inference. To model the structure of CM incidence and forecast the future infection of time series, it is essential to build an effective model aiming to improve forecasting performance.

Deep learning has been widely considered or utilized to address the challenges related to forecasting models in various domains/topics, including emerging infectious disease prediction. As one of the most popular deep learning algorithms, Long Short-Term Memory (LSTM) is a special case of recurrent neural network (RNN) initially introduced by Hochreiter and Schmidhuber.[27] LSTM offers advantages over traditional regression methods, especially when dealing with sequential or time-dependent data. For example, LSTM can capture both short-term and long-term dependencies within data sequences, model complex, nonlinear relationships between variables and is better suited for handling missing data, noise and irregularities.[27] In addition, deep learning models including LSTM do not need to validate assumptions and define hypotheses in advance compared to statistical methods.[27] LSTMs have shown the strength to be highly effective in a variety of time series forecasting problems with high accuracy, such as stock market financial forecasting,[28] petroleum production forecasting,[29] and predictive modeling in public health including dengue incidence forecasting,[30] influenza forecasting,[31] and epidemic transmission forecasting.[32] In response to COVID-19, several LSTM models were built to perform epidemic spread prediction, such as COVID-19 time series forecasting.[33,34]

For the empirical analysis, we used the estimated daily level pneumonia due to CM, as derived from a large urgent care service in Maricopa County, Arizona from 2020 to 2022.[35] Since our study period falls within the COVID-19 pandemic, in addition to widely used weather and air quality data, we also consider the influence of the restrictions on people's daily activities. Studies found that human activities may play a more influential role than climate in explaining anomalies in CM prevalence.[36] The restrictions and prevention measures, coupled with public fear of contagion during the COVID-19 pandemic, have resulted in an unparalleled transformation of people's daily activity.[37] Many researchers found that the number and types of out-of-home activities changed significantly during the COVID-19 pandemic.[38] However, it remains unknown whether and how the change in people's daily activities affected cocci infection. To get a better understanding of how the restrictions and people's daily activity changes were related to CM infection, we included a stringency index in the model.

In this study, we apply LSTM to forecast the future trend of CM incidence. Instead of arbitrarily selecting the input variables, we use statistical methods, including cross-correlation, collinearity tests, and stepwise regression, to guide the selection process. This study is the first to introduce LSTM as a new method of forecasting CM infection in time series with accurate trend prediction. Also, the CM community can benefit from this study to assist public health effectively.

## Methods

### Daily estimated prevalence of new CM in Maricopa County

Our study focuses on Maricopa County, Arizona which accounts for approximately one-half of all US CM cases. Since CM is often undiagnosed because appropriate testing is not performed, we used pneumonia cases tested for CM that were found to be positive to estimate the proportion of all urgent care pneumonia patients that likely had CM. In this research, we used clinical data from Banner Urgent Care Services (BUCS) to estimate the prevalence of new coccidioidal infections in this community.[35] Although BUCS cases of CM constitute less than 3% of all cases reported to Maricopa County Department of Public Health, the monthly correlation of BUCS and County cases is very strong, ($r = 0.86$). Briefly, the number of BUCS patients diagnosed with CAP (pneumonia, organism unspecified, ICD10 code = J18), those tested for CM, and the percentage of their CM tests that were positive were collected daily from BUCS' electronic medical record. The percentage of tested patients that were positive for CM was multiplied by the total number of patients diagnosed with CAP to estimate the total number or CAP patients with CM. The University of Arizona Institutional Review Board has determined that this activity does not constitute human experimentation.

For the period 2020 through 2022, there were 10,945 patients diagnosed with CAP. Since the daily frequency of CAP was relatively low, a simple 30-day moving average was calculated to determine the estimated daily CAP patients with CM between 2020 and 2022. This was calculated by summing the number of events in a particular day and the prior 29 days divided by 30. Calculating the moving averages can also mitigate the random noise in the dataset. Fig. 1 presents the temporal changes of the confirmed CM cases by year. Of note, the frequency of CAP in the first half of 2020 may have included patients with COVID-19 prior to the availability of a diagnostic test for that disease, but subsequently patients found to have COVID-19 were categorized with a different diagnostic code than CAP. Shown on Fig. 1, the temporal variation of CM from 2020 to 2022 is complex, making it difficult to identify consistent seasonal or monthly patterns in the confirmed cases. Traditional regression models, which assume linear or simple nonlinear relationships between variables, often perform poorly on nonlinear, non-stationary data, leading to suboptimal modeling and

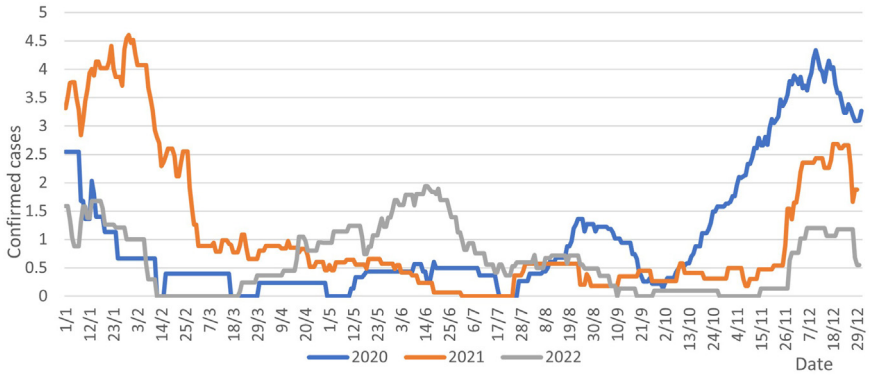


***Fig. 1:*** 30-day moving average of confirmed CM cases from 2020 to 2022. Blue, orange and green lines represent the temporary variation of CM cases respectively. From the figure, we can hardly identify the seasonal/monthly pattern of the CM confirmed cases.

prediction of CM infections.[23,24] Additionally, the absence of clear seasonal or monthly patterns complicates the selection of features that account for seasonality. Thus, traditional statistical methods may compromise the prediction performance, making LSTM a more suitable method for this case study.

### Data and data pre-processing

For the explanatory variables, the weather and precipitation data for this analysis were drawn from publicly available data sets collected and maintained by the Maricopa County Flood Control District. The weather variables come from the Weather Sensor Data.[39] Data is collected daily at 40 real-time weather stations. Variables collected include mean temperature, maximum temperature, minimum temperature, mean dew point, maximum dew point, minimum dew point, maximum relative humidity, minimum relative humidity, maximum peak wind, maximum air pressure, minimum air pressure, and maximum solar radiation. Considering that *Coccidioides* grows in the soil of endemic regions, spores are less likely to emerge within urban areas.[20] As a result, we selected a subset of 14 weather stations in non-urban areas of Maricopa County (Supplementary Figure S1). We aggregated the Weather Sensor Data to create a countywide daily average by taking the average value for a variable per day across the selected sensors.

The precipitation data comes from the Historic Precipitation dataset.[39] This data is recorded daily at 357 precipitation gages throughout Maricopa County and the surrounding counties. A subset of this data was taken, focusing only on the 193 sensors in non-urban areas of Maricopa County (Supplementary Figure S2). This data was also aggregated to create a countywide daily average by taking the average across the selected sensors each day.

Air quality data for this analysis comes from the Environmental Protection Agency's (EPA) Outdoor Air Quality Data.[40] This data is collected and maintained by the EPA Air Quality System. For this analysis we look at the concentration of Particulate Matter 10 ($PM_{10}$) and Particulate Matter 2.5 ($PM_{2.5}$). $PM_{10}$ and $PM_{2.5}$ refer, respectively, to inhalable particles that generally have a diameter of 10 μm or smaller and generally have a diameter of 2.5 μm or smaller. Both metrics are measured in terms of mass per cubic meter of air. This data is measured at 21 different sites in Maricopa County, which we aggregated by taking a daily average across all sensors.

Like with the counts of confirmed CM cases, a simple 30-day moving average is calculated for the weather, precipitation, and air quality data. We also consider the cumulative effects of humidity and wind speed by incorporating metrics for continuous days of high relative humidity (CDHH) and continuous days of high wind speed (CDHW) into the model. High relative humidity and high wind speed are defined as conditions exceeding the mean plus one standard deviation. CDHH and CDHW quantify the number of consecutive days that meet these criteria for relative humidity and wind speed, respectively.

Drought data was collected from the U.S. Drought Monitor,[41] which is publicly available and maintained by the National Drought Mitigation Center at the University of Nebraska–Lincoln. Data is reported weekly. Our analysis utilized the Drought Severity and Coverage Index (DSCI) for Maricopa County as the metric for drought. This variable is created by the U.S. Drought Monitor as a way of expressing drought levels for a whole area in a single variable. It considers how much of a geographic area is experiencing each level of drought. A DSCI value of 0 means that none of the areas are abnormally dry or in drought. A DSCI of 500 means that all the areas are experiencing exceptional drought. To normalize the data, we used a 4-week moving average for this metric. For the weather, precipitation, air quality, and drought

variables we looked at data from January 2020 through December 2022.

The stringency index of Maricopa County is calculated by the Oxford Coronavirus Government Response Tracker (OxCGRT) project.[42] The stringency index is a composite measure based on several response indicators such as school closures, workplace closures, and travel bans, scaled to a value from 0 to 100 (100 = strictest). Similar to CM confirmed cases, we also calculate the 30-day moving average of the potential influence factor.

## LSTM

LSTM network models were first introduced by Hochreiter and Schmidhuber.[27] The LSTM models use memory cells to capture long-term dependencies and have a set of gates to regulate the flow of information. These key features allow LSTMs to effectively model sequences and temporal patterns, making them particularly well-suited for tasks such as time-series forecasting.[32]

Fig. 2[43] gives a visual representation of an LSTM cell. The LSTM model consists of an input layer, hidden layers, and a fully connected layer at the end. Each hidden layer consists of a defined number of units, which defines the dimensionality of the output space. Specifically, the input gate $i_t$, output gate $O_t$, forget gates $f_t$ and activation function *Act Func* are used to model LSTMs and learn the behavior of temporal correlations. The input gate controls if the incoming data should be input to the cell. The output gate controls if the passing data should be in the output hidden state. The forget gate controls if the data should be removed from the cell. A hidden layer also contains a memory cell, which retains important information and discards unnecessary data from previous inputs during training.

In our empirical analysis, during the training phase, CM infection forecasts were generated using a combination of historical input data, that includes both observed CM infection counts and the selected independent variables from previous training iteration(s), along with the independent variables from the current iteration. The forecasting process continues as the time window shifts forward. During the testing phase, however, the approach differs: instead of using observed CM case counts as prior input data, the model utilized the predicted values from earlier time steps.

Activation functions are necessary to form mappings between an input and output in neural network model layers.[44] If these mappings are non-linear in nature, they are easier to optimize and can handle more complex interpretations of data,[45] which is necessary while building neural networks. Three activation functions that have been used in epidemic prediction research[46] are considered in this work, which are sigmoid, tanh, and rectified linear unit (ReLU). By testing different

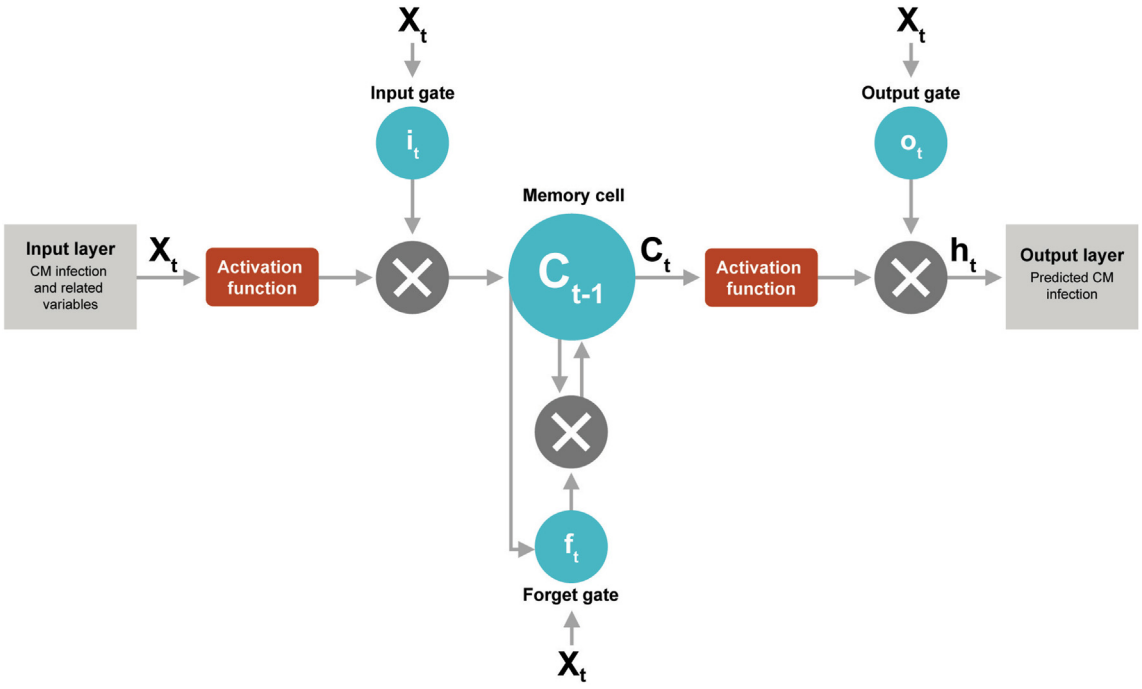


*Fig. 2:* Architecture of an LSTM Cell. The LSTM model consists of an input layer, hidden layers, and a fully connected layer at the end. In our study, $X_t$ represents the CM infection counts and the independent variables from previous training iteration(s), along with the independent variables from the current iteration.

activations, we can finally select a LSTM network that can effectively handle the dataset in our case, balancing the flow of information and maintaining stability during training. The general formulation of each function is presented below, where x is the weighted combination of the current input and the previous hidden state, plus a bias term to the layer.

The sigmoid function (output range from 0 to 1) is mathematically defined as below:

$$f(x) = \frac{1}{1 + e^{-x}}$$

The tanh function (output range from –1 to 1) is shaped similarly to the sigmoid but is symmetric around the origin. It is mathematically defined as:

$$f(x) = (e^{x} - e^{-x}) \Big/ (e^{x} + e^{-x})$$

The ReLU function (output range from 0 to infinity) is mathematically defined as:

$$f(x) = \max(0, x)$$

We also added dropout[47] to each of the hidden layers. During training, the dropout technique randomly removed a percentage of units from a layer to prevent overfitting. For instance, a dropout rate of 0.2 would mean that 20% of the units in that layer are randomly dropped. The final layer in the model is a fully connected dense layer, which specifies the number of future days to forecast. For example, if the model predicts 30 days in the future, the number of units in the dense layer would be 30.

The optimizer used in our model is Adam since it is the most popular optimizer known for its computational efficiency and memory optimization.[48,49] The loss function for our model is the mean squared error (MSE). MSE is used as the measure to evaluate the model performance in our study because it is a reliable indicator of how close the test and predicted values are.

Optimizing the parameters of a neural network is important to maximize its performance. Determining what value to set to the parameters ends up being empirical in nature. In our work, we cycled through different values for the batch size, dropout, and activation function. Building a model with the most accurate predictions is an empirical task that relies on fine-tuning the parameters. We implemented models with different parameter values and monitored the results for each combination. This helps us determine the best combinations of parameters and values. Table 1 details the different parameter values we experimented with. The LSTM analysis was conducted using Python 3.9.8. The code will be available upon request.

| Parameter | Values |
|---|---|
| Batch size | 2, 5, 10 |
| Dropout | 0.1, 0.2, 0.3 |
| Activation function | ReLU, tanh, sigmoid |

*Table 1:* **The parameters and their corresponding values.**


### Role of the funding source

The funders had no role in study design, data collection, data analysis, interpretation, or writing of the report.


## Results

### Variable selection

Statistical analyses were performed using SPSS 28.0.1.1 (15) to determine which variable should be included in the LSTM model and to understand the relationship between weather, climate variables, and stringency index with CM infection. First, we conducted a cross-correlation analysis between each of the independent variables and the confirmed pneumonia cases. Cross-correlation analysis is a statistical analysis method allowing us to estimate the degree of association between two waveforms. By conducting the cross-correlation analysis, we can get not only the correlation between the potential influential factors and the infection cases but also the time lags of the correlation. Table 2 presents the variables with a statistically significant correlation with CM infection. Temperature, relative humidity (RH), wind speed, precipitation, CDHH, and CDHW have a negative association with confirmed pneumonia cases with a time lag ranging from 12 to 56. For example, a temperature increase will lead to a decrease in CM infection in 12–17 days. While other factors, such as dew point, air pressure, and solar radiation are not correlated with CM infection.

Collinearity between the independent variables can compromise the performance of the prediction model. Considering the time lag effect, we examined the correlation between each two variables to diagnose the potential collinearity problem (Table 3). From Table 3 we can find that relative humidity has a relatively high correlation with other variables, especially temperature, precipitation, $PM_{10}$, and continued days of high relative

| Variable | Correlation | Time lag |
|---|---|---|
| Mean temperature | Negative | 12–17 days ahead |
| Relative humidity (RH) | Negative | 50–56 days ahead |
| Wind speed | Negative | 35–38 days ahead |
| CDHH | Negative | 35–37 days ahead |
| CDHW | Negative | 38–39 days ahead |
| Precipitation | Negative | 35–38 days ahead |
| $PM_{10}$ concentration | Positive | Simultaneously |
| $PM_{2.5}$ concentration | Positive | Simultaneously |
| Drought | Positive | Simultaneously |
| Stringency | Positive | Simultaneously |

*Table 2:* **Cross-correlation analysis results.**

| | Temp | RH | Wind speed | Precip | $PM_{10}$ | $PM_{2.5}$ | CDHW | CDHH | Drought | Stringency |
|---|---|---|---|---|---|---|---|---|---|---|
| Temp | 1 | −0.428[b] | −0.183[b] | 0.100[b] | 0.199[b] | −0.465[b] | −0.270[b] | −0.252[b] | 0.145[b] | 0.142[b] |
| RH | −0.428[b] | 1 | 0.163[b] | 0.650[b] | −0.590[b] | 0.093[b] | 0.184[b] | 0.473[b] | −0.379[b] | −0.357[b] |
| Wind speed | −0.183[b] | 0.163[b] | 1 | −0.084[b] | 0.156[b] | −0.241[b] | 0.695[b] | 0.443[b] | 0.112[b] | −0.198[b] |
| Precip | 0.100[b] | .650[b] | −0.084[b] | 1 | −0.519[b] | −0.243[b] | −0.092[b] | 0.318[b] | −0.154[b] | −0.117[b] |
| $PM_{10}$ | 0.199[b] | −0.590[b] | 0.156[b] | −0.519[b] | 1 | 0.354[b] | 0.073[a] | −0.166[b] | 0.520[b] | 0.147[b] |
| $PM_{2.5}$ | −0.465[b] | 0.093[b] | −0.241[b] | −0.243[b] | 0.354[b] | 1 | −0.073[a] | −0.041 | 0.147[b] | −0.074[a] |
| CDHW | −0.270[b] | 0.184[b] | 0.695[b] | −0.092[b] | 0.073[a] | −0.073[a] | 1 | 0.654[b] | 0.056 | −0.155[b] |
| CDHH | −0.252[b] | 0.473[b] | 0.443[b] | 0.318[b] | −0.166[b] | −0.041 | 0.654[b] | 1 | −0.119[b] | −0.245[b] |
| Drought | 0.145[b] | −0.379[b] | 0.112[b] | −0.154[b] | 0.520[b] | 0.147[b] | 0.056 | −0.119[b] | 1 | 0.163[b] |
| Stringency | 0.142[b] | −0.357[b] | −0.198[b] | −0.117[b] | 0.147[b] | −0.074[a] | −0.155[b] | −0.245[b] | 0.163[b] | 1 |

[a]Significant at 0.10 level ($p < 0.10$). [b]Significant at 0.05 level ($p < 0.05$).

*Table 3:* **Collinearity test.**

humidity. As a result, we excluded relative humidity in the LSTM model implementation.

We used the stepwise regression method to further select variables that were highly correlated with CM infection and removed the variables whose variation could be represented by other variables or the combination of other variables. Finally, variables, including temperature, precipitation, wind speed, $PM_{10}$ concentration, $PM_{2.5}$ concentration, drought, and stringency index showed a unique influence on CM infection and were included in the LSTM analysis.

## LSTM prediction

After excluding the period with continuous zero confirmed cases, we used the first 990 days, from January 1st, 2020, to September 17th, 2022, to run the LSTM analysis. Training on a dataset that contains continuous zeros is not advisable as it does not yield meaningful forecasts. We predicted the daily confirmed pneumonia cases, considering normalized measures for temperature, precipitation, wind speed, $PM_{10}$ concentration, stringency index, drought, and $PM_{2.5}$ concentration.

Table 4 details the different window sizes and forecast combinations we implemented. For example, taking 30 days as input, we trained the model to forecast for the next 10 days, 20 days, and 30 days, respectively. Similarly, given 90 days as input, the next 30 days were forecasted, and so on.

For each combination of window size and forecast period, we evaluated various parameter values as outlined in Table 1, resulting in 567 scenarios. We used three methods to identify the optimal combinations. First, we employed visual judgment by reviewing forecast graphs to assess the fit between actual and predicted curves for both the training and forecast periods, focusing on trends, peaks, valleys, and synchronicity. Next, we calculated the validation loss, which measures the error on the validation dataset after each training epoch and helps monitor generalization and detect overfitting.[50] Additionally, we assessed the accuracy of each scenario using the MSE metric.

We selected two combinations with different lengths of forecasting periods (shown in Table 5). In the first model, we used 120 days of historical data to predict the coming 10 days of CM confirmed cases (Fig. 3). In the second model, the confirmed cases in the coming 30 days were predicted based on 120 days of historical data (Fig. 4). We found a tradeoff between the length of the forecasting period and the performance of the forecasting result. The shorter forecasting period has a lower MSE, indicating a better model performance. The MSE for the first model is relatively low at 0.0039 when compared to the MSE of the second model at 0.011. While using 120 days of historical data to forecast the next 30 days of confirmed CM cases can predict the trend of CM infection fairly well, the predicted values are less accurate compared to those for the next 10 days.

| Window size | Forecast |
|---|---|
| 30 | 10 |
| | 20 |
| | 30 |
| 60, 90, 120 | 10 |
| | 20 |
| | 30 |
| | 40 |
| | 50 |
| | 60 |

*Table 4:* **Combinations of window sizes and forecasts.**

| ID | Combination | Parameter value | | | MSE |
|---|---|---|---|---|---|
| | (Window size_forecast) | Dropout | Activation function | Batch size | |
| Model 1 | 120_10 | 0.1 | sigmoid | 5 | 0.0039 |
| Model 2 | 120_30 | 0.2 | tanh | 2 | 0.011 |

*Table 5:* **The MSE and parameters for the recommended combinations.**

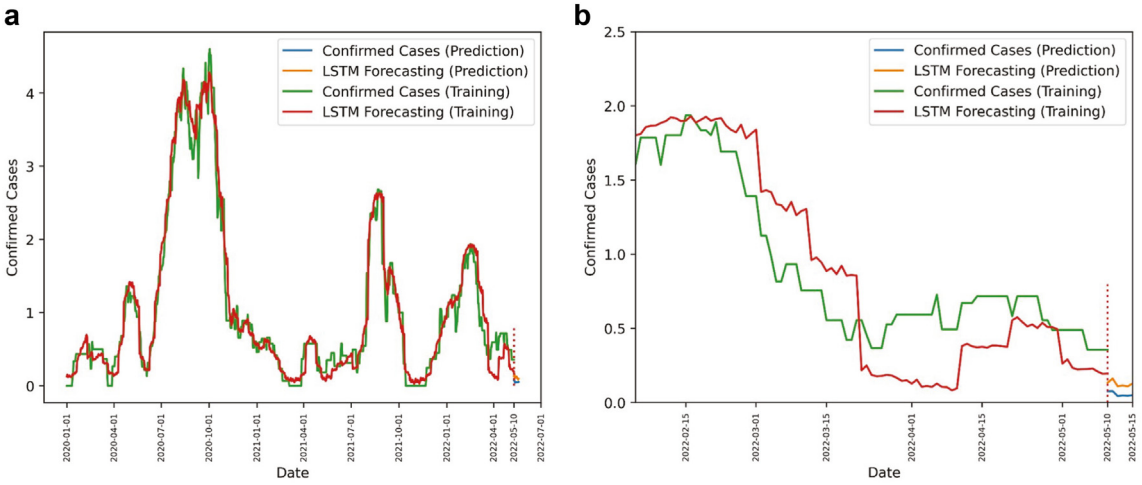


***Fig. 3:*** LSTM Forecasting Result for Model 1. Fig. 3a presents the prediction results for the entire dataset; Fig. 3b zooms to the prediction part. Blue and green line represent the actual confirmed cases; the red line represents the forecasting performance using the training dataset and the orange line is the forecasting result of the prediction period.

Fig. 3 depicts the model's performance with the selected parameters for the recommended combination Model 1. Fig. 3a presents the prediction result for the entire dataset and Fig. 3b zooms to the prediction part. Fig. 4 highlights the performance of Model 2. In both figures, the green and blue lines represent the actual confirmed CM cases, with the green line showing the cases used for model training and the blue line serving as a reference for comparison with the model's predictions. The red line displays the LSTM model's performance based on the training data, and the orange line depicts the forecasted outcomes. Fig. 5 depicts the training and validation loss graphs for Model 1 and Model 2, respectively. During training, the model fits well to both upward and downward trends, with validation loss approaching zero by the end of the epoch run. The LSTM forecast closely aligns with the original trend during prediction, with an average difference of 0.06 for Model 1, and 0.1 for Model 2. All variables included in the LSTMs are significant in terms of feature importance measurement (Supplementary Figures S3 and S4).

## Discussion

This is the first attempt to utilize deep learning in CM research, specifically for prediction of the CM incidence rates. As shown in our recent study of CM in Maricopa County,[35] the seasonality of high and low incidence of CM varies from year to year, and past

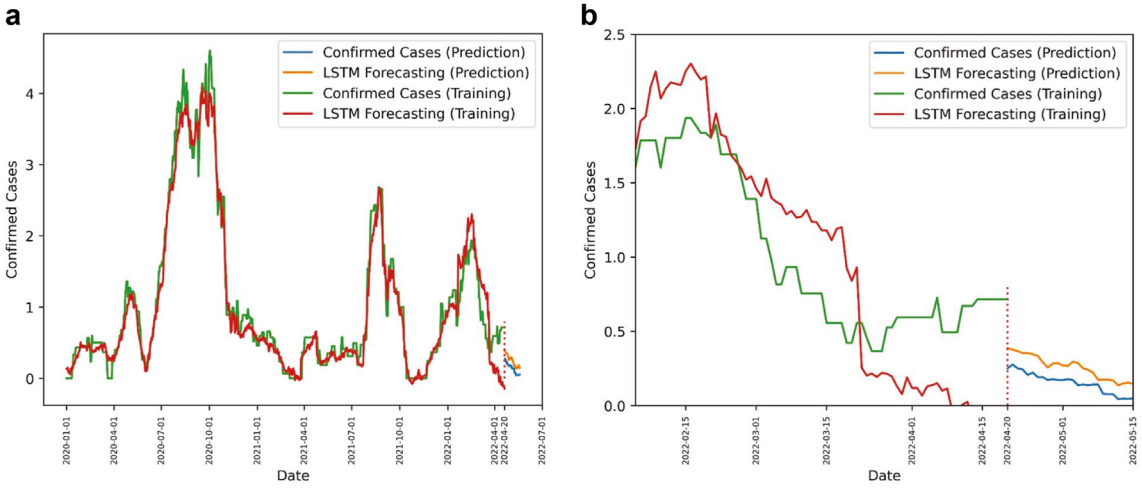


***Fig. 4:*** LSTM Forecasting Result for Model 2. Fig. 4a presents the prediction results for the entire dataset; Fig. 4b zooms to the prediction part. Blue and green line represent the actual confirmed cases; the red line represents the forecasting performance using the training dataset and the orange line is the forecasting result of the prediction period.

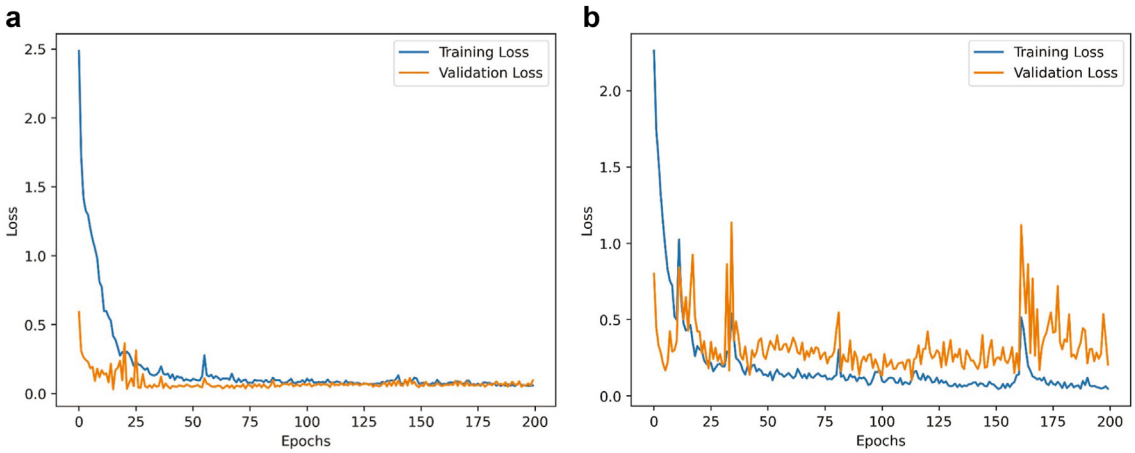


*Fig. 5:* Loss Graph of Training vs Validation. Fig. 5a depicts the training and validation loss graphs for Model 1 and Fig. 5b represents the result for Model 2. Both of the validation loss approaching zero by the end of the epoch run.

studies have indicated that this is strongly influenced by weather patterns.[4] Having accessible predictions of how CM incidences will behave in the coming weeks would be very useful for clinicians in deciding how aggressively to test for CM and for public health agencies to provide advice to the community at risk. Given the nature of deep learning, it is essential to train an LSTM model iteratively by finetuning input variables, model parameters, and training/testing data. It is also necessary to train a developed LSTM model again when incorporating new data, such as the most recent estimated CM infection counts, to improve the model learning process. The performance of LSTM models can vary in terms of the selections of variables, model parameter values, and data by researchers or clinicians. However, as the foundation of the analysis, the deep learning algorithm LSTM is the same and the algorithm implementation process is consistent. This is the fundamental component of the methodology which can benefit the CM research community.

In this study, we ran a series of statistical analyses between the potential factors and the CM infection count to select the input variables instead of having an arbitrary selection. Besides the widely considered

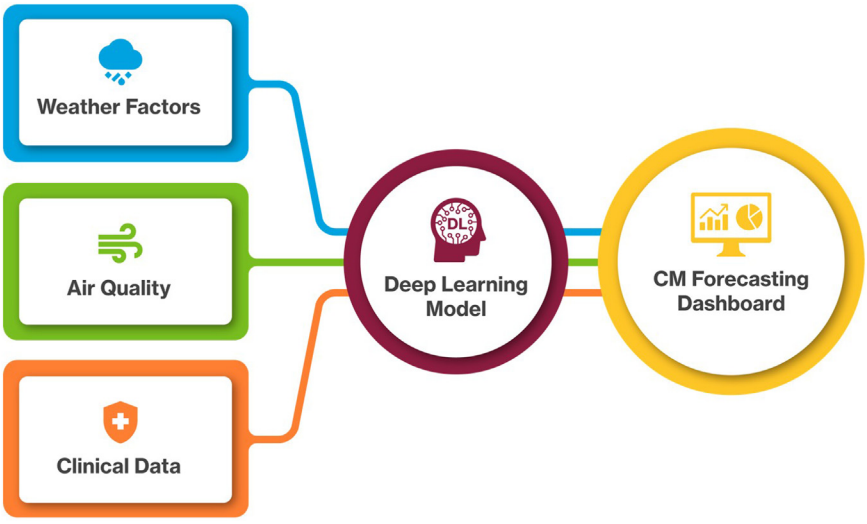


*Fig. 6:* Future framework for CM forecasting. The framework consists of three components: (1) real-time data collection through APIs for air quality, weather, and clinical data to enable forecasting on larger, up-to-date datasets; (2) continuous refinement of deep learning models, such as AutoEncoders and CNN Encoder-LSTM Decoders, by retraining on recent datasets to improve predictive accuracy; and (3) development of an interactive CM forecasting dashboard to provide decision support for clinicians, including ER applications, using a user-friendly interface with visual aids like color-coded schemes to guide testing and improve public health outcomes.

weather and climate factors, we also included the stringency index to reflect the restrictions and constraints on people's out-of-home activities, as our study period covers different stages of COVID-19. After removing the factors which were highly correlated with other influential factors, we included temperature, precipitation, wind speed, $PM_{10}$ concentration, $PM_{2.5}$ concentration, drought, and stringency index in LSTM algorithm. The results showed that LSTM can be applied in CM prediction with accurate trend.

In practical applications, LSTM models can be built to forecast the CM infection with different lengths of prediction periods. It is worth mentioning the tradeoff between the prediction period and the model performance. The model with a longer prediction period has a relatively higher MSE. Overall, LSTM models can help the research community in CM conduct and even improve prediction from a new analytical perspective. More importantly, they can help researchers, clinicians, and the public have a better understanding of CM prediction to support public health.

One important limitation of this study is the clinical data from BUCS only covers a small percentage of all CM cases reported to Maricopa County Department of Public Health. Underreporting of valley fever cases[6,7] remains a potential issue, as not all cases are captured in the dataset. This narrower focus might impact the forecasting results. Furthermore, the dataset includes several continuous days with zero reported cases, a factor that can compromise the overall performance of the LSTM by introducing periods of low variability. These limitations highlight the need for broader and more comprehensive datasets in future research to improve model accuracy and applicability.

Future work needs to analyze the spatial variation of CM and predict the trend both temporally and spatially. It would be worthwhile to improve the prediction performance by incorporating more relevant variables into the model, such as soil factors, socio-demographic factors, and people's travel and activity behaviors. On the other hand, it would also be valuable to simplify LSTM models with fewer common factors (e.g. weather and air quality) for a general use to make an LSTM model more portable. A broad group of CM researchers and clinicians could use the simplified model to conduct prediction analysis with minimum algorithm implementation. Additionally, the model can be trained on a larger dataset to gain wider insights into the general CM trend. Besides improving the accuracy of the model, future work should also address the effects of climate on CM patterns. In theory, these effects could include an expansion of the range of CM or shifting patterns within the endemic region. The consequences of both types of effects could be dramatic, but further research is needed to see if and how climate change can affect CM.

Specifically, a framework for future work could be developed to enhance the study's utility and scalability, consisting of three main components (as illustrated in Fig. 6). Firstly, real-time data collection can be facilitated using Application Programming Interfaces (APIs) for air quality, weather factors, and clinical data, enabling forecasting on larger and more recent datasets. Secondly, deep learning models should be continuously updated and refined by retraining on these extensive, up-to-date datasets, thereby improving the model's learning process in both training and testing phases. Advances in cutting-edge deep learning models, such as AutoEncoders and CNN Encoder LSTM Decoder architectures, provide promising opportunities for future CM forecasting. Lastly, creating an interactive CM forecasting dashboard would allow researchers and clinicians to perform their own CM forecasting, making these sophisticated models more accessible and practical for widespread use. One direct application of these types of models is that they can, once validated, be incorporated into easy-to-use dashboards that can provide decision support for clinicians, especially in ER settings where time is of the essence. Turning the predictions of the model into a simple color scheme that suggests when to run tests for CM will improve public health and efficiency in clinical settings.

#### Contributors

All authors made substantial contributions to the article.

Conception and design: SSK and FW.
Literature review: XJ, FW, SSK, and KS.
Development of methodology: FW, SSK, XJ, and TU.
Acquisition of data: XJ, KS, SSK, and JNG.
Analysis and interpretation of data: XJ, FW, SSK, and TU.
Verification of data and results: XJ, FW, SSK, and TU.
Writing, review, and/or revision of the manuscript: All authors.
Supervision: SSK, FW, and MDL.
Funding acquisition: SSK and JNG.

All authors had access to the data in the study. FW, SSK, JNG, and MDL had final responsibility for the decision to submit for publication.

#### Data sharing statement

The clinical data supporting this study was obtained from Banner Health System. The data is de-identified and will only be available upon request. All other raw data used in this study are publicly available and links are provided in the references.

#### Declaration of interests

We declare no competing interests.

#### Acknowledgements

Research reported in this manuscript was supported by the grant, "Getting to the Source of Arizona's Valley Fever Problem: A Tri-University Collaboration to Map and Characterize the Pathogen Where It Grows", from the Arizona Board of Regents.

#### Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.lana.2025.101010.

#### References

1 Galgiani JN, Kauffman CA. Coccidioidomycosis and histoplasmosis in immunocompetent persons. *N Engl J Med*. 2024;390(6):536–547.
2 Dobos RR, Benedict K, Jackson BR, McCotter OZ. Using soil survey data to model potential Coccidioides soil habitat and inform Valley fever epidemiology. *PLoS One*. 2021;16(2):e0247263.

3 Gorris ME, Treseder KK, Zender CS, Randerson JT. Expansion of coccidioidomycosis endemic regions in the United States in response to climate change. *Geohealth*. 2019;3(10):308–327.
4 Comrie AC. Climate factors influencing coccidioidomycosis seasonality and outbreaks. *Environ Health Perspect*. 2005;113(6):688–692.
5 Coopersmith EJ, Bell JE, Benedict K, Shriber J, McCotter O, Cosh MH. Relating coccidioidomycosis (valley fever) incidence to soil moisture conditions. *GeoHealth*. 2017;1(1):51–63.
6 Benedict K, Ireland M, Weinberg MP, et al. Enhanced surveillance for coccidioidomycosis, 14 US states, 2016. *Emerg Infect Dis*. 2018;24(8):1444.
7 National Academies of Sciences, Engineering, and Medicine. *Impact and Control of Valley Fever: Proceedings of a Workshop–in Brief*. 2023.
8 Kahn D, Chen W, Linden Y, et al. A microbial risk assessor's guide to Valley Fever (Coccidioides spp.): case study and review of risk factors. *Sci Total Environ*. 2024;917:170141.
9 California department of public health. Available from: https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/ValleyFeverDashboard.aspx; 2024.
10 Arizona department of health services. Available from: https://www.azdhs.gov/preparedness/epidemiology-disease-control/valley-fever/index.php#reports-publications; 2024.
11 Tamerius JD, Comrie AC. Coccidioidomycosis incidence in Arizona predicted by seasonal precipitation. *PLoS One*. 2011;6(6):e21009.
12 Weaver EA, Kolivras KN. Investigating the relationship between climate and valley fever (coccidioidomycosis). *EcoHealth*. 2018;15:840–852.
13 Meisner J, Clifford WR, Wohrle RD, Kangiser D, Rabinowitz P. Soil and climactic predictors of canine coccidioidomycosis seroprevalence in Washington State: an ecological cross-sectional study. *Transbound Emerg Dis*. 2019;66(5):2134–2142.
14 Ocampo-Chavira P, Eaton-Gonzalez R, Riquelme M. Of mice and fungi: Coccidioides spp. distribution models. *J Fungi*. 2020;6(4):320.
15 Weaver E, Kolivras KN, Thomas RQ, Thomas VA, Abbas KM. Environmental factors affecting ecological niche of Coccidioides species and spatial dynamics of valley fever in the United States. *Spat Spatiotemporal Epidemiol*. 2020;32:100317.
16 Head JR, Sondermeyer-Cooksey G, Heaney AK, et al. Effects of precipitation, heat, and drought on incidence and expansion of coccidioidomycosis in western USA: a longitudinal surveillance study. *Lancet Planet Health*. 2022;6(10):e793–e803.
17 Kollath DR, Teixeira MM, Funke A, Miller KJ, Barker BM. Investigating the role of animal burrows on the ecology and distribution of Coccidioides spp. in Arizona soils. *Mycopathologia*. 2020;185:145–159.
18 Gorris ME, Cat LA, Zender CS, Treseder KK, Randerson JT. Coccidioidomycosis dynamics in relation to climate in the southwestern United States. *GeoHealth*. 2018;2(1):6–24.
19 Lauer A, Etyemezian V, Nikolich G, et al. Valley fever: environmental risk factors and exposure pathways deduced from field measurements in California. *Int J Environ Res Publ Health*. 2020;17(15):5285.
20 Brown HE, Wangshu M, Mohammed K, Clarisse T, Jian L, Daoqin T. Spatial scale in environmental risk mapping: a Valley fever case study. *J Public Health Res*. 2017;6(2):886.
21 Siami-Namini S, Tavakoli N, Namin AS. A comparison of ARIMA and LSTM in forecasting time series. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE; 2018:1394–1401.
22 NASA. Extreme weather and climate change. Available from: https://science.nasa.gov/climate-change/extreme-weather/; 2023.
23 Nelson BK. Time series analysis using autoregressive integrated moving average (ARIMA) models. *Acad Emerg Med*. 1998;5(7):739–744.
24 Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*. 2003;50:159–175.
25 Hidalgo B, Goodman M. Multivariate or multivariable regression? *Am J Publ Health*. 2013;103(1):39–40.
26 Mayr A, Fenske N, Hofner B, Kneib T, Schmid M. Generalized additive models for location, scale and shape for high dimensional data—a flexible approach based on boosting. *J Roy Stat Soc C Appl Stat*. 2012;61(3):403–427.
27 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–1780.
28 Cao J, Li Z, Li J. Financial time series forecasting model based on CEEMDAN and LSTM. *Phys Stat Mech Appl*. 2019;519:127–139.
29 Sagheer A, Kotb M. Time series forecasting of petroleum production using deep LSTM recurrent networks. *Neurocomputing*. 2019;323:203–213.
30 Mussumeci E, Coelho FC. Machine-learning forecasting for dengue epidemics-comparing LSTM, random forest and lasso regression. *medRxiv*. 2020:2020–2021.
31 Zhu H, Chen S, Lu W, et al. Study on the influence of meteorological factors on influenza in different regions and predictions based on an LSTM algorithm. *BMC Publ Health*. 2022;22(1):2335.
32 Barman A. Time series analysis and forecasting of covid-19 cases using LSTM and ARIMA models. *arXiv*. 2020. https://doi.org/10.48550/arXiv.2006.13852.
33 Azhie A, Sharma D, Sheth P, et al. A deep learning framework for personalised dynamic diagnosis of graft fibrosis after liver transplantation: a retrospective, single Canadian centre, longitudinal study. *Lancet Digit Health*. 2023;5(7):e458–e466.
34 Abbasimehr H, Paki R, Bahrini A. A novel approach based on combining deep learning models with statistical methods for COVID-19 time series forecasting. *Neural Comput Appl*. 2022;34(4):3135–3149.
35 Galgiani JN, Lang A, Howard BJ, et al. Access to urgent care practices improves understanding and management of endemic coccidioidomycosis: Maricopa county, Arizona, 2018–2023. *Am J Med*. 2024;137:951 [accepted].
36 Zender CS, Talamantes J. Climate controls on valley fever incidence in Kern County, California. *Int J Biometeorol*. 2006;50:174–182.
37 Anwari N, Ahmed MT, Islam MR, Hadiuzzaman M, Amin S. Exploring the travel behavior changes caused by the COVID-19 crisis: a case study for a developing country. *Transp Res Interdiscip Perspect*. 2021;9:100334.
38 De Vos J. The effect of COVID-19 and subsequent social distancing on travel behavior. *Transp Res Interdiscip Perspect*. 2020;5:100121.
39 Flood Control District of Maricopa County. Weather sensor data. Available from: https://www.maricopa.gov/3769/Weather-Sensor-Data; 2024.
40 United States environmental protection agency. Available from: https://www.epa.gov/outdoor-air-quality-data/download-daily-data; 2023.
41 National Integrated Drought Information System. *U.S. Drought Monitor (USDM)*; 2000. Available from: https://www.drought.gov/data-maps-tools/us-drought-monitor.
42 Mathieu E, Ritchie H, Rodés-Guirao L, et al. Coronavirus pandemic (COVID-19). Published online at OurWorldInData.org. Available from: https://ourworldindata.org/coronavirus; 2020.
43 Shastri S, Singh K, Kumar S, Kour P, Mansotra V. Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study. *Chaos, Solit Fractals*. 2020;140:110227.
44 Sibi P, Jones SA, Siddarth P. Analysis of different activation functions using back propagation neural networks. *J Theor Appl Inf Technol*. 2013;47(3):1264–1268.
45 Chung H, Lee SJ, Park JG. Deep neural network using trainable activation functions. In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE; 2016:348–352.
46 Pal R, Sekh AA, Kar S, Prasad DK. Neural network based country wise risk prediction of COVID-19. *Appl Sci*. 2020;10(18):6448.
47 Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15(1):1929–1958.
48 Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv*. 2014. https://doi.org/10.48550/arXiv.1412.6980.
49 Desai C. Comparative analysis of optimizers in deep neural networks. *Int J Innovat Sci Res Technol*. 2020;5(10):959–962.
50 Salman S, Liu X. Overfitting mechanism and avoidance in deep neural networks. *arXiv*. 2019. https://doi.org/10.48550/arXiv.1901.06566.