Online-Journals.org

iTDAF IETI Transactions on Data Analysis and Forecasting

iTDAF | eISSN: 2959-0442 | Vol. 1 No. 2 (2023) | OPEN ACCESS

https://doi.org/10.3991/itdaf.v1i2.40175

PAPER

Early Prediction of Monkeypox Virus Outbreak Using Machine Learning

Solomon Oluwole Akinola[1](✉), Qing-Guo Wang[1,2], Peter Olukanmi[1], Tshilidzi Marwala[1]

[1]Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa

[2]BNU-HKBU United International College, Zhuhai, China

oluwolea@uj.ac.za

ABSTRACT

At the onset of an infectious disease, such as the monkeypox virus (MPXV), surveillance data is crucial in keeping track of the outbreak's progression. The surveillance data for MPXV received considerable attention after multiple European countries recorded cases. Historical data obtained from May 9, 2022, to August 10, 2022, were used to model the cumulative case trajectories of MPXV in five countries. Our study employed autoregressive integrated moving averages (ARIMA), neural network autoregression (NNETAR), exponential smoothing (ETS), and seasonal naïve regression (SNAÏVE) for training and evaluation. The paper makes the following contributions: (1) enhanced model stability with the Box-Cox transformation as a preprocessing step, (2) experimentation with both linear and non-linear models, and (3) simulation of the top five countries during the impulsive rise in cases of MPXV. The results were evaluated using three metrics: root mean square error (RMSE), mean square error (MAE), and mean absolute percentage error (MAPE). The ARIMA (0,1,3) (1,0,0)[7] model yielded the lowest percentage error of 5.16 in the holdout set for MAPE in France observations. The ETS (A, A, A) model, the lowest percentage error in the holdout set for MAE was 7.35 in Germany. Regarding the NNETAR (1,1,2) [7] model, the lowest percentage error in the holdout observations for RMSE was 8.33 in Spain, 2.75 in the United Kingdom (UK), and 8.05 in the United States of America (USA) in that order. Based on these findings, we can conclude that while the transformation proved crucial for model performance, it was not necessary for all experiments, as ARIMA remained dominant in France and the ETS model in Germany. At the same time, NNETAR model outperformed in cumulative case counts in Spain, the UK, and the USA. Our experimentation allows for early identification and contributes to a better understanding of forecasting MPXV cases using combinations of both linear and nonlinear models.

KEYWORDS

monkeypox virus, epidemiological analysis, time series, auto-regression, exponential smoothing, and neural network

Akinola, S.O., Wang, Q.-G., Olukanmi, P., Marwala, T. (2023). Early Prediction of Monkeypox Virus Outbreak Using Machine Learning. *IETI Transactions on Data Analysis and Forecasting (iTDAF)*, 1(2), pp. 14–29. https://doi.org/10.3991/itdaf.v1i2.40175

Article submitted 2023-04-05. Resubmitted 2023-05-15. Final acceptance 2023-05-16. Final version published as submitted by the authors.

NOMENCLATURE

AIC	Akaike Information Criterion
A.R.	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
ETS	Exponential Smoothing
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MPXV	Monkeypox Virus
RMSE	Root Mean Square Error
USA	United States of America
UK	United Kingdom
WHO	World Health Organization

1 INTRODUCTION

Epidemiological analysis plays a central role in the public health system of a country [1]. Healthcare-associated infections and microbial threats pose significant health risks [2]. The emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [3] and the recent increase in cases of monkeypox virus (MPXV) cases in countries where outbreaks are rare or non-existent [4] raises concern. Time-series techniques allows for forecasting of future dynamics, enabling the development of robust surveillance systems to characterize MPXV diseases in a population [4],[5]. Epidemiological time-series analyses utilize past surveillance data to forecast future disease incidence. Researchers have applied various time-series models in related studies using statistical [6], machine learning, deep learning, and hybridized models [1]; [5]–[7]. Regarding MPXV, forecasting cases is challenging due to the limited availability of surveillance data. In one of the earliest reports [8], cumulative case growth was compared between two periods, while a recent study used timely insights early human judgment forecasts [9]. Alternatively, the classical approach can offer governments and organizations by forecasting MPXV cases using surveillance data.

MPXV is a zoonotic virus that belongs to the orthopoxvirus family [8]. Martín-Delgado et al., [10] noted that the first such case occurred in the early seventies. MPXV was previously known to be endemic in Central African countries [4], [11]. In 2003, multiple cases were reported in the USA among individuals in contact with animals imported from outside the country [12]. More recently, MPXV cases emerged in Portugal, Spain, and the UK in May 2022 [11]. On July 23, 2022, the World Health Organization (WHO) declared MPXV as a global health emergency [13] based on surveillance data from six WHO regions reporting suspected cases. The report further noted that transmission occurs through close and intimate physical contact with infected individuals. The WHO recommends control measures, including vaccination [11], as well as preventive measures, including infection regulation in non-human primates such as rodents, squirrels, and dormice [10].

Various techniques have been employed for epidemiological forecasting, using both linear and non-linear techniques. When modelling MXPV transmission, the susceptible, exposed, infectious, and recovered (SEIR) framework has been

considered [14]. Linear approaches, such as autoregressive integrated moving average (ARIMA) [15] and exponential smoothing (ETS) [16], are commonly used as generalized baseline approaches in several forecasting tasks. In recent years, non-linear machine learning approaches have gained popularity. One technique used in infectious disease modelling is the incorporation of the artificial neural network (ANN) with other generalized time-series techniques, as demonstrated by [17]. The selection of features is crucial in determining models for epidemiological time-series forecasting, including factors such as incidence rate, region, population diversity, and health status. Recent forecast approaches have utilized hybrid models that combine both linear and non-linear techniques characterized with comparable levels of accuracies and forecast horizons [18]–[22].

Based on historical data, we developed short-term forecasts for MPXV cases in five countries. Furthermore, a comparative analysis of three assessment metrics, the root mean square error (RMSE), mean square error (MAE), and mean absolute percentage error (MAPE), was conducted to evaluate the performance of the four models. There was a strong correlation between the MPXV cumulative case test data and the results obtained from other models. The study reveals that NNETAR outperformed all the other techniques in four out of the five countries. The NNETAR results were applied to all three-performance metrics. ARIMA outperformed the other techniques in one of the five countries, and this superiority was consistent across all three metrics. We further observed that NNETAR outperformed ARIMA in three countries, while ARIMA performed better than others techniques for the cumulative MPXV case counts in the USA. Our analysis suggests that the efforts to reduce MPXV spread in the four European countries were part of a unified strategy adopted across the European states. In contrast, case counts emerged in the USA several weeks after the European countries had reported cases. The measures implemented in the USA to control MPXV spread did not align with the observed case counts during the time of the experiment. In Germany, after applying the Box-Cox transformation, the ETS technique demonstrated superior performance compared to other techniques. The SNAÏVE regression approach was implemented as a benchmark.

This study conducted a comparative analysis of time-series forecasting for MPXV surveillance data from the five most impacted countries as of August 10, 2022. Epidemiological time-series analysis is well-established, encompassing both linear and non-linear models. In this study, ARIMA, ETS, and NNETAR techniques were employed to forecast MPXV cases, showcasing their potentials for early MPXV predictions. The remainder of this paper is organized as follows. Section II discusses the surveillance data, techniques, and performance metrics used in the study. Section III present the details of the model experiments and identifies the most suitable model for MPXV surveillance data. Finally, the conclusions are presented in Section IV.

2 METHOD

2.1 Data

The MPXV surveillance data used in this study were obtained from the freely shared Our World in Data repository [12]. Table I lists the countries in the MPXV dataset along with their corresponding cumulative case counts. For evaluation

and forecasting purposes, the cumulative daily case counts of the five countries were considered up until August 10, 2022. The data observations were modest, and a higher-ranking forecasting technique was expected as MPXV surveillance data evolved. Approximately 30,000 MPXV cases have been reported. Eighty-seven countries without historical MPXV cases account for 92%, and seven countries with historical MPXV cases only accounting for 8% [23]. The explanatory variables were the daily time-interval stamps, while the response variables were the cumulative case counts of MPXV surveillance data in the five most-impacted countries under investigation. The surveillance record of France and Germany's spanned a duration of eighty-four days, starting from May 19, 2022. In Spain and the USA, surveillance records began on May 18, 2022, and lasted for eighty-five days. In the UK, it started on May 6, 2022, and covered a period of ninety-four days.

2.2 Data transformation

We applied the Box–Cox transformation to the MPXV using surveillance data to increase the distribution of residuals and stabilize the variance [24], as indicated by the value of λ in Table 1. The transformation was implemented [25], which is relevant for selecting variance-stabilizing and bias-reduction techniques. The resulting series from the Box-Cox transformation exhibited a uniform Gaussian distribution. However, it should be noted that, assumptions of normality were violated due to heavily tailed data observations, resulting in less informative regression analysis. Nevertheless, the Box–Cox transformation ensured stability in the power transformation. Density distribution plots from Figures 1 to 5 were presented for the five most-impacted countries prior to the application of the Box-Cox transformation. These distribution plots were right-tailed distributions and failed the critical test for an approximately normal distribution. Visual observation were combined with statistical tests to demonstrate acceptable deviations from the standard lines. The Box-Cox transformation then applied to significantly boost the model fit.

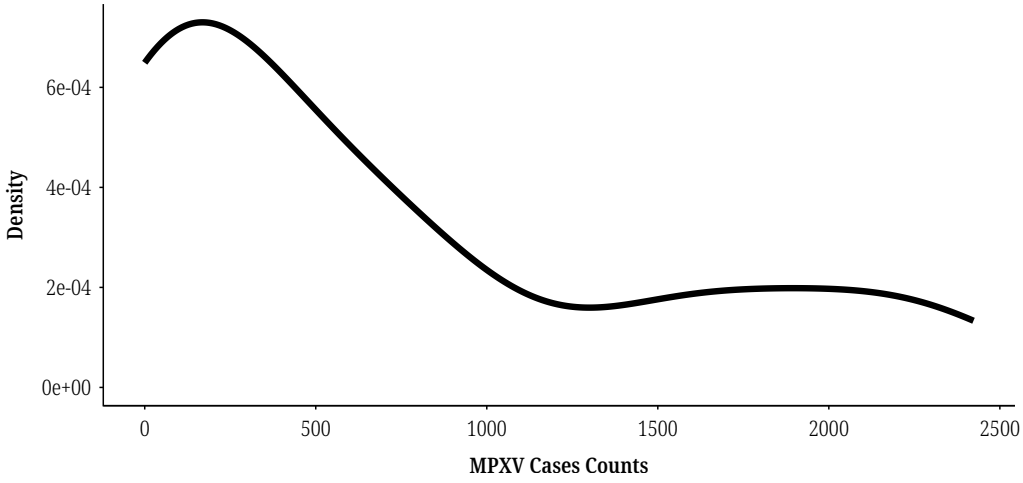

Fig. 1. MPXV Density distribution in France from June 19th to August 10th 2022

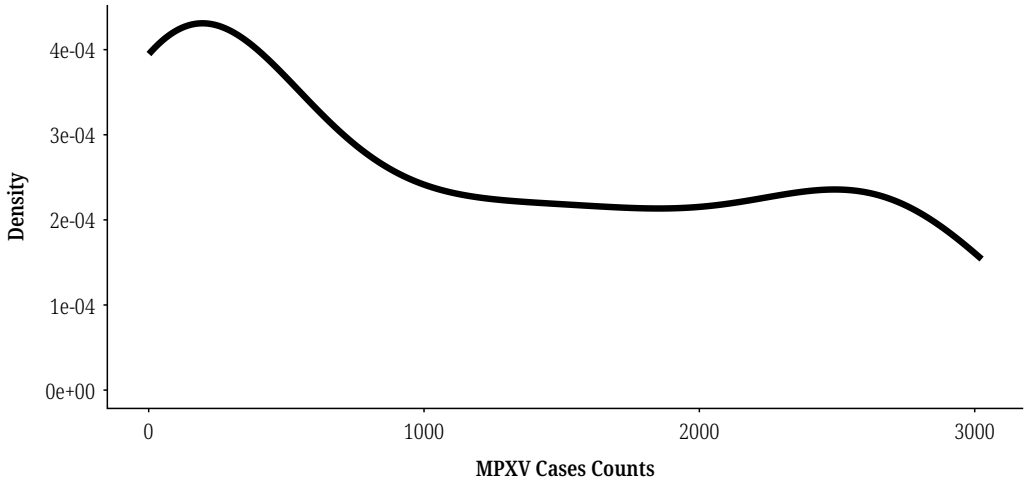

Fig. 2. MPXV Density distribution in Germany from May 20th to August 10th 2022

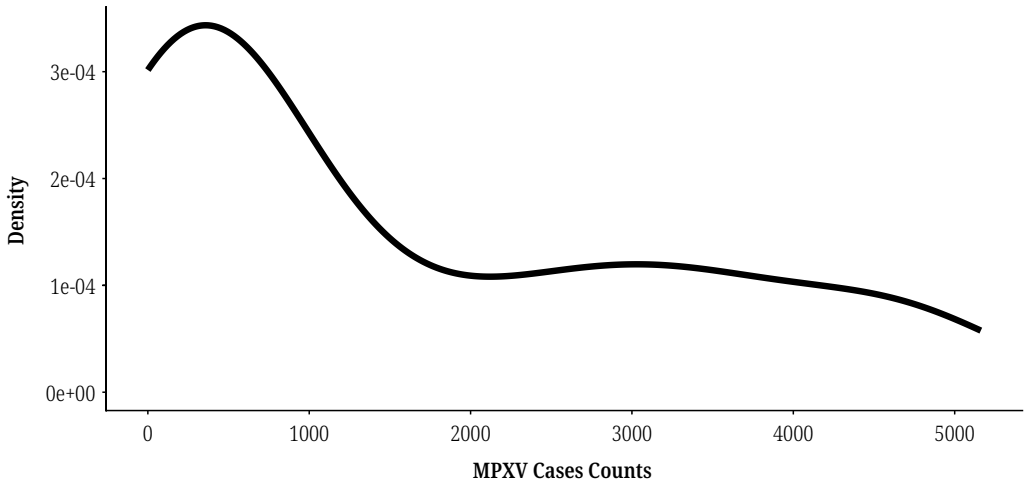

Fig. 3. MPXV Density distribution in Spain form May 18th to August 10th 2022

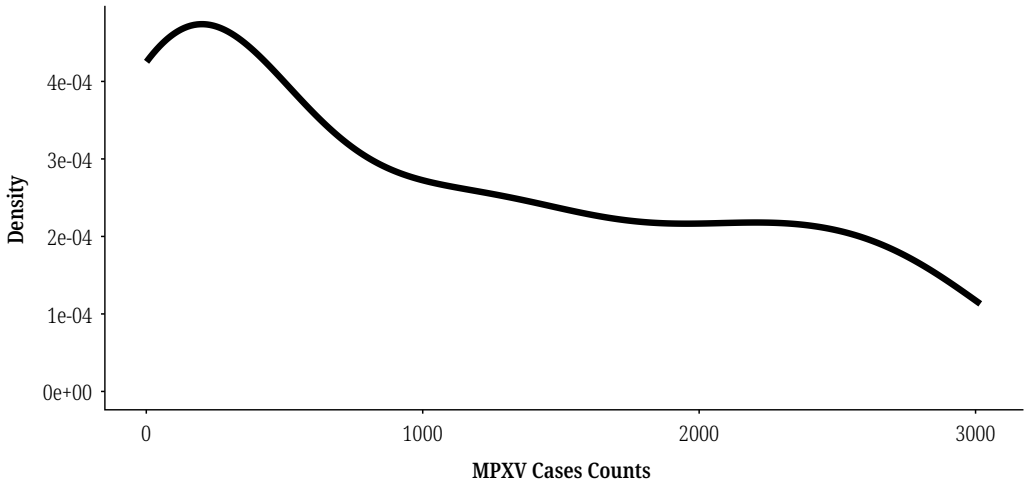

Fig. 4. MPXV Density distribution in the UK from May 7th to August 8th 2022

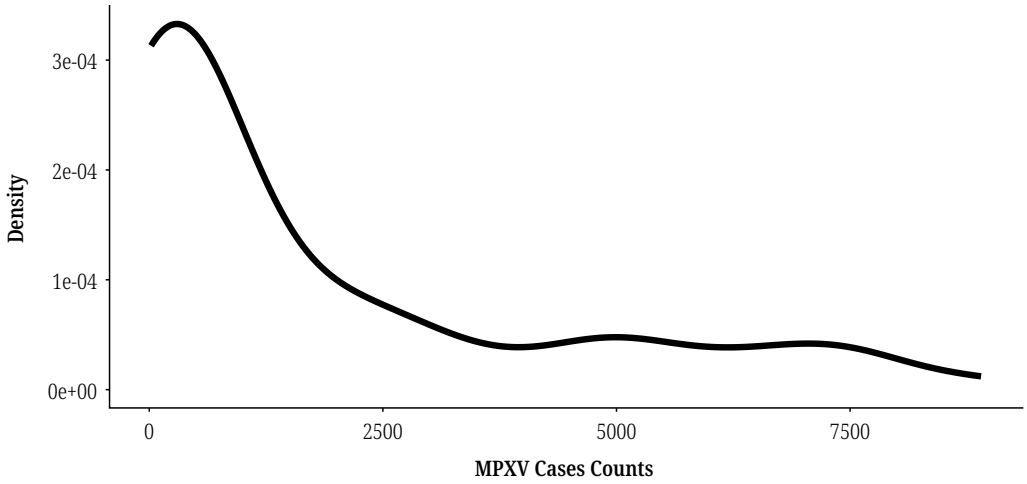

Fig. 5. MPXV Density distribution in the USA from June 6th to August 10th 2022

$$x_i^{(\lambda)} = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ ln(x_i) & \text{if } \lambda = 0. \end{cases} \tag{1}$$

In equation (1), λ is the index log transformation, and the observation time x_i is nonzero for the observed time-series. The transformed shifts for France, Germany, Spain, the UK, and the USA are listed in Table 1. France had the smallest converted λ value, whereas Germany had the highest transformed value.

Table 1. MPXV cases and transformed values

S/N	Country	Cumulative Cases	Λ*
1	France	2591	0.42
2	Germany	3025	0.51
3	Spain	5162	0.34
4	United Kingdom	3022	0.48
5	USA	10360	0.04

Note: *MPXV case Box-Cox values as λ.

2.3 MPXV surveillance data decomposition

The time-series exhibits multiple compositions [26]. In the MPXV surveillance data for the five countries, each underlying component contributes significantly to the time series, including the three underlying components: trend, seasonal, and remainder.

2.4 Methodology

The **Naïve Method** was popularized for financial time series [26], and it was proposed that variations in observations are similar in distribution and independent of

each other. In the naive technique, future forecasts are based on previous observations, as illustrated in equation (2).

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}, \tag{2}$$

where $\hat{y}_{T+h|T}$ is the estimation of $y_1,\ldots,y_T$, m is the seasonal period, and k is the number of completed periods before a forecast.

The **Autoregressive Integrated Moving Average (ARIMA)**, developed by Box and Jenkins [15], is a well-established time series analysis model. In equation (3), three components are estimated for ARIMA (p, d, q).

$$\hat{y}_t = c + \phi_1 \hat{y}_{t+1} + \ldots + \phi_p \hat{y}_{p+1} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-p} + \varepsilon_t. \tag{3}$$

The $\hat{y}_{t+n}$ forecast is a function of multiple orders of p, autoregressive (A.R.) component, d, order of difference, and q, order of moving average (M.A.). c is the mean difference between successive observations. The ε_t component specifies the model white noise, ϕ_1, ϕ_2, ..., ϕ_p, is the A.R. component, and θ_1, θ_2, ..., θ_p, represents the M.A. components.

The **Neural Network Autoregression (NNETAR)** method combines neural networks with augmented reality (A.R.) parameters. Neural networks comprise three layers: input, hidden, and output [26]. A node in a layer is connected in two parts: the summation of hidden inputs with weights and an activation function to determine the summed output. Neural networks are forecasting techniques used for non-linear observations. NNETAR is a hybrid model capable of modelling complex linear and non-linear relationships. NNETAR (p, o) has two components: p lags and o hidden-layer nodes. In seasonal NNETAR (p, P, o) m, P is the number of seasonal lags, and m is the seasonal component.

The **Exponential Smoothing (ETS)** model utilizes weighted lags that decay exponentially [16]. The unobserved components of ETS, including error, trend, and seasonality, are described using state-space models. ETS is widely applied in the financial industry where similar variable observations are encountered. ETS models encompass combinations of none (N), additive (A), and multiplicative (M) components for errors, trends, and seasonal states. For example, ETS (A, N, N) is a simple exponential smoothing with an additive error, ETS (M, N, N) is a simple exponential smoothing with a multiplicative error, and other state-space exponential smoothing techniques can be derived. ARIMA, NNETAR, and ETS models have been well evaluated in epidemiological time-series forecasting, with competitive evaluation results in test data.

2.5 Evaluation

Mean Square Error (MAE) represents the mean scale of the inaccuracy from a set of forecasts. Specifically, *MAE* determines the variation between absolute and forecast values [27]. Equation (4) illustrates *MAE*.

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|\hat{y}_t - y_t\right|. \tag{4}$$

Root Mean Square Error (RMSE) is an error metric for the average scale of errors. *RMSE* describes how concentrated the data fit the optimal model. The RMSE

is a widely applicable metric for time-series forecasting tasks. Equation (5) illustrates the RMSE.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}{T}}. \tag{5}$$

Mean Absolute Percentage Error (MAPE) can also be described as mean absolute percentage deviation (MAPD). The *MAPE* denotes the significance of the error values relative to the observed values. Equation (6) describes *MAPE* as

$$MAPE = \frac{\sum_{t=1}^{T}\left(\frac{y_t - \hat{y}_t}{y_t}\right)}{T}. \tag{6}$$

In (4), (5), and (6), $\hat{y}_t$ denotes the predicted value, y_t denotes the actual value, and T is the number of observations.

2.6 Workflow

As shown in Figure 6, the MPXV cumulative case data contained no missing data. We split the MPXV surveillance data into training and test sets. The Box-Cox technique was used to transform the training data, which were used to train ARIMA, ETS, NNETAR, and SNAIVE models. The models were evaluated using the MPXV cumulative case test data.

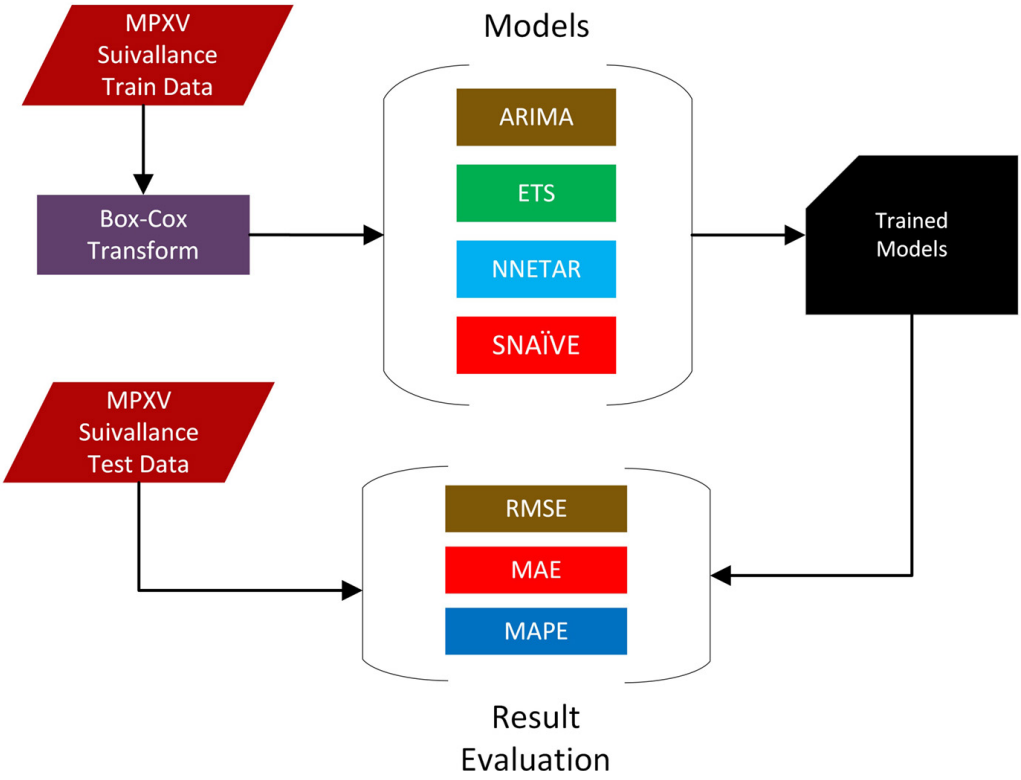

Fig. 6. Proposed workflow for the MPXV cumulative case-forecasting task

3 RESULTS AND DISCUSSION

The models for MPXV surveillance data from the five countries were developed using the R Studio, along with the fpp3 [26], and timetk packages [28]. The training dataset consisted of 91% of the MPXV surveillance data, while 9% was reserved for testing. The independent MPXV surveillance test data for each country were evaluated using the ARIMA, NNETAR, SNAIVE, and ETS models. In the ARIMA model selection, the Auto.Arima fpp3 package was used to optimize the discovery of the Akaike information criterion (AIC) approximation for model estimation. NNETAR offers competitive estimations by automatically determining the appropriate specifications for lags and neurons using the fpp3 package. The best-performing ETS model was automatically selected from the optimized AIC discovery using the fpp3 package. The test data results were obtained using the ARIMA, NNETAR, SNAIVE, and ETS models.

We conducted a seven-days forecast for the five most-impacted countries starting from August 4, 2022. The results of the forecasting models are presented in Tables 2 and 3, and the corresponding plots are shown in Figures 7 to 11. To evaluate the performance of the models, we used the RMSE, MAE, and MAPE metrics. The NNETAR model demonstrated superior performance in Spain and the UK without applying the Box-Cox transformation. However, after implementing the Box-Cox transformation, the model's performance improved in the USA, resulting in a total of three countries where NNERAR outperformed. On the other hand, the ARIMA model outperformed in France, Germany, and the USA before applying the Box-Cox transformation. Implementing the Box-Cox transformation failed to improve the error scores in France and Germany.

Table 2. Performance evaluation of MPXV surveillance data

Model	RMSE	MAE	MAPE	Country
NNETAR (1,1,2) [7]	3.38E+02	3.28E+02	1.40E+01	France
ETS(A,Ad,A)	4.39E+01	3.68E+01	1.59E+00	
ARIMA (0,2,1) (0,0,1) [7]	**1.63E+02**	**1.56E+02**	**6.72E+00**	
SNAIVE	4.03E+02	4.02E+02	1.74E+01	
NNETAR (1,1,2) [7]	1.22E+02	1.16E+02	3.95E+00	Germany
ETS(A,A,A)	5.01E+01	4.49E+01	1.55E+00	
ARIMA(1,2,4)(1,0,0)[7]	**4.84E+01**	**4.53E+01**	**1.55E+00**	
SNAIVE	2.79E+02	2.78E+02	9.54E+00	
NNETAR(1,1,2)[7]	**2.40E+02**	**1.93E+02**	**3.88E+00**	Spain
ETS(A,A,N)	2.72E+02	2.46E+02	5.20E+00	
ARIMA(0,2,1)	2.60E+02	2.32E+02	4.92E+00	
SNAIVE	7.48E+02	7.39E+02	1.56E+01	
NNETAR (1,1,2) [7]	**4.61E+01**	**3.95E+01**	**1.36E+00**	United Kingdom
ETS(A,A,A)	1.06E+02	1.03E+02	3.61E+00	
ARIMA (1,2,2)	1.00E+02	9.09E+01	3.22E+00	
SNAIVE	3.02E+02	2.99E+02	1.06E+01	
NNETAR (1,1,2) [7]	1.68E+03	1.51E+03	1.98E+01	USA
ETS(M,A,N)	4.16E+02	3.34E+02	4.33E+00	
ARIMA (0,2,1) (1,0,0) [7]	**2.80E+02**	**2.42E+02**	**3.28E+00**	
SNAIVE	2.53E+03	2.50E+03	3.42E+01	

Table 3. Performance evaluation of MPXV surveillance data with box-cox transformation

Model	RMSE	MAE	MAPE	Country
NNETAR (1,1,2) [7]	2.49E+02	2.39E+02	1.02E+01	France
ETS(A,A,A)	9.71E+01	8.98E+01	3.89E+00	
ARIMA (0,1,3)(1,0,0)[7] w/ drift	**9.24E+01**	**7.61E+01**	**3.25E+00**	
SNAIVE	3.70E+02	3.69E+02	1.59E+01	
NNETAR (1,1,2) [7]	7.25E+01	6.60E+01	2.25E+00	Germany
ETS(A,A,A)	**5.04E+01**	**4.16E+01**	**1.44E+00**	
ARIMA (2,1,2)(1,0,1)[7] w/ drift	5.62E+01	4.89E+01	1.67E+00	
SNAIVE	2.53E+02	2.52E+02	8.64E+00	
NNETAR (1,1,2) [7]	**2.20E+02**	**1.72E+02**	**3.45E+00**	Spain
ETS(A,A,N)	5.77E+02	5.43E+02	1.13E+01	
ARIMA (0,1,0) w/ drift	5.81E+02	5.47E+02	1.13E+01	
SNAIVE	6.50E+02	6.39E+02	1.35E+01	
NNETAR (1,1,2) [7]	**3.34E+01**	**3.20E+01**	**1.12E+00**	United Kingdom
ETS(A,A,A)	1.60E+02	1.48E+02	5.17E+00	
ARIMA (0,1,1) (1,0,0) [7] w/ drift	1.56E+02	1.50E+02	5.24E+00	
SNAIVE	2.80E+02	2.77E+02	9.82E+00	
NNETAR (1,1,2) [7]	**3.27E+02**	**2.79E+02**	**3.68E+00**	USA
ETS(A,A,A)	1.40E+03	1.29E+03	1.71E+01	
ARIMA(0,1,1)(2,0,0)[7] w/ drift	9.88E+02	8.40E+02	1.10E+01	
SNAIVE	1.69E+03	1.66E+03	2.26E+01	

Implementation of the Box-Cox transformation resulted in lower errors when compared to models without transformation, except for the ETS(A, A, A) model for France, Germany, the UK, and the USA; and the ETS(A, A, N) model for Spain. Similar results were observed for the ARIMA (2,1,2)(1,0,1)[7] model w/drift for Germany, ARIMA (0,1,0) w/drift for Spain, ARIMA (0,1,1) (1,0,0) [7] w/drift for the UK, and ARIMA (0,1,1)(2,0,0)[7] w/drift model for the USA, except for ARIMA (0,1,3)(1,0,0)[7] w/drift for France. The drift is an equivalent alignment between the start and end point observations, leading to forecast extrapolation [26]. The percentage changes in RMSE for the ARIMA model were 76%, –72%, –36%, –55%, and –14% for France, the USA, the UK, Spain, and Germany, respectively. In the same country order, the 105%, –71%, –39%, –58%, and –7% for MAE and 107%, –70%, –39%, –56%, and –7% MAPE was observed. The percentage changes in RMSE for the ETS model were –55%, –1%, –53%, –34%, and –70% for France, Germany, Spain, the UK, and the USA, respectively. Moreso, in the same country order, –59%, 8%, –55%, –30%, and –74% for MAE and –59%, 8%, –54%, –30%, and –75% for MAPE was registered. The percentage changes in RMSE for the NNETAR model were 36%, 68%, 9%, 38%, and 414% for France, Germany, Spain, the UK, and the USA, respectively. Again, in the same country order, 37%, 76%, 12%, 23%, and 441% for MAE and 37%, 76%, 12%, 21%, and 438% for MAPE was noted. The percentage changes in RMSE for the SNAIVE model were 9%, 10%, 15%, 8%, and 50% for France, Germany, Spain, the UK, and the USA, respectively. In the same country order, for MAE and MAPE, it was found to be 9%, 10%, 16%, 8%, and 51%, respectively.

In the results of the Box-Cox transformation for the five most-impacted countries, the France ARIMA (0,2,1) (0,0,1) [7] model parameters had no A.R. component, a second lag order difference in the nonseasonal model, an A.R. order of two, and an M.A. order of one. In contrast, the seasonal component had an M.A. lag order of one. In the Germany ARIMA(1,2,4)(1,0,0)[7] model parameter, the A.R. had a lag order of one, a second lag order difference, and an M.A. lag order of four and one in the nonseasonal and seasonal components, respectively. The Spain ARIMA(0,2,1) model parameter had a nonseasonal second-order lag difference and an M.A. lag order of one. The UK ARIMA (1,2,2) had a nonseasonal A.R. with lag one, a second lag order difference, and an M.A. lag order of two. ARIMA (0,2,1) (1,0,0) [7] parameters, a nonseasonal second lag order difference and a nonseasonal M.A. lag order of one and one A.R. lag in the seasonal component, respectively.

NNETAR had one input lag in the first nonseasonal value, one seasonal input value, two neurons in the hidden layer, and seven neurons, as indicated by the seasonal period for all the five most-impacted countries by August 10, 2022. The five most-affected countries in the ETS(A, A, A) model were purely addictive.

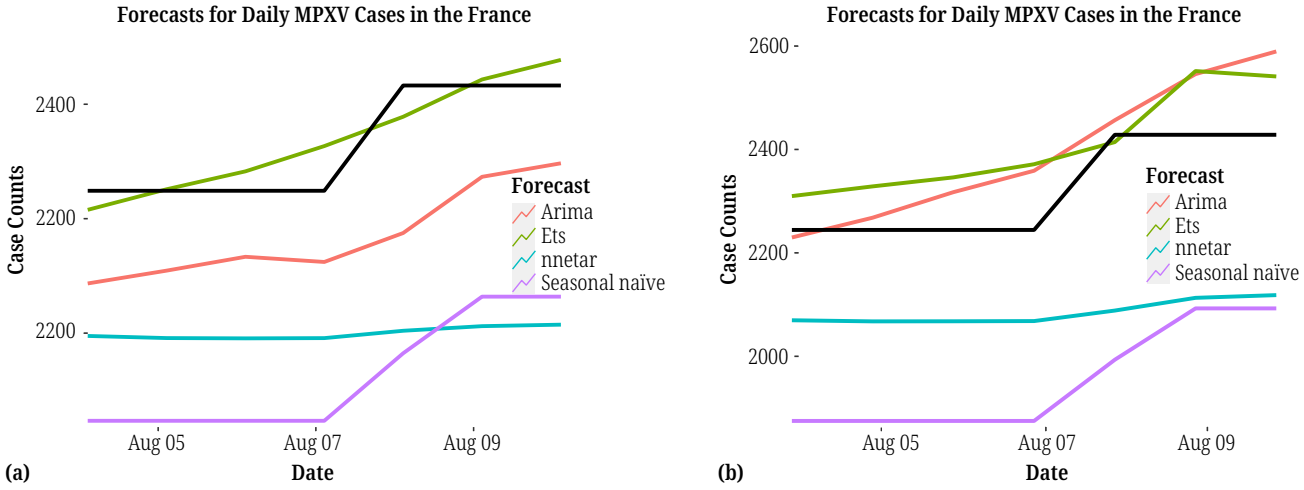

Fig. 7. France's seven-day forecast from August 4, 2022, to August 10, 2022: (a) without Box-Cox transformation and (b) with Box-Cox transformation

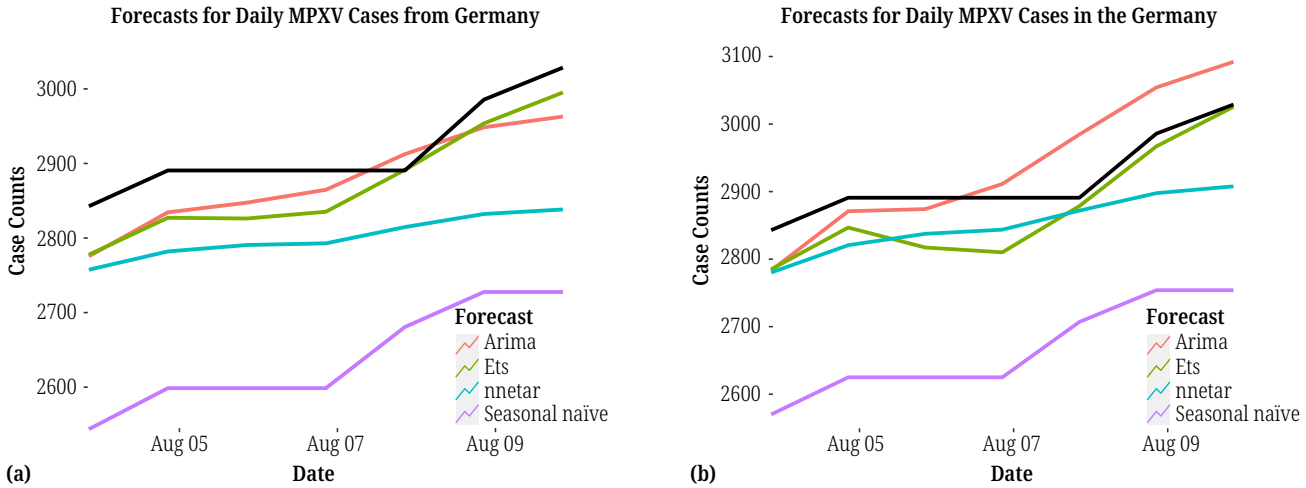

Fig. 8. Germany's seven-day forecast from August 4, 2022, to August 10, 2022: (a) without Box-Cox transformation and (b) with Box-Cox transformation

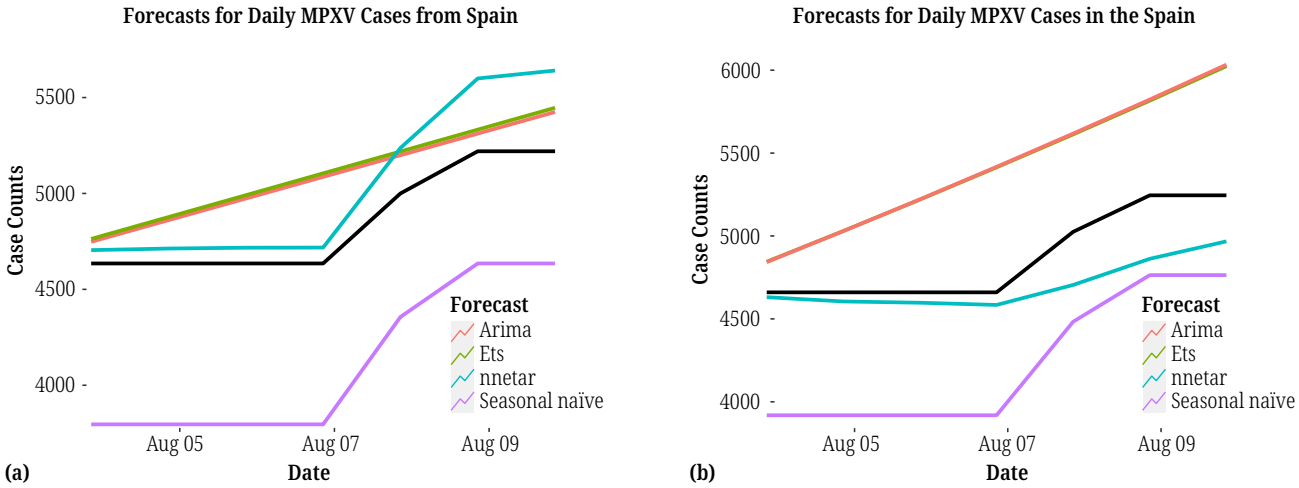

Fig. 9. Spain's seven-day forecast from August 4, 2022, to August 10, 2022: (a) without Box-Cox transformation and (b) with Box-Cox transformation

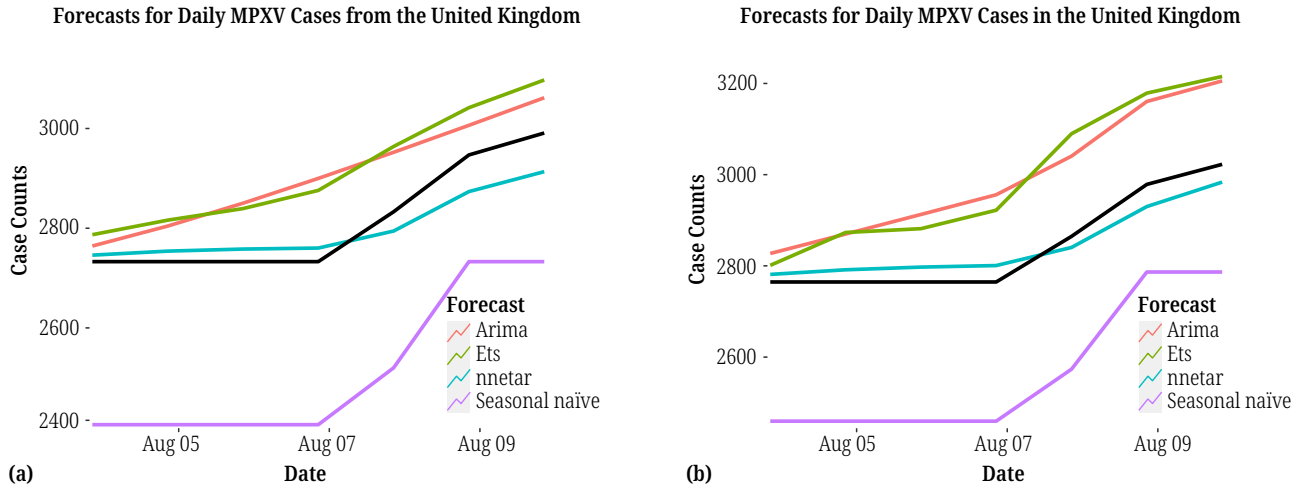

Fig. 10. The United Kingdom's seven-day forecast from August 4, 2022, to August 10, 2022: (a) without Box-Cox transformation and (b) with Box-Cox transformation

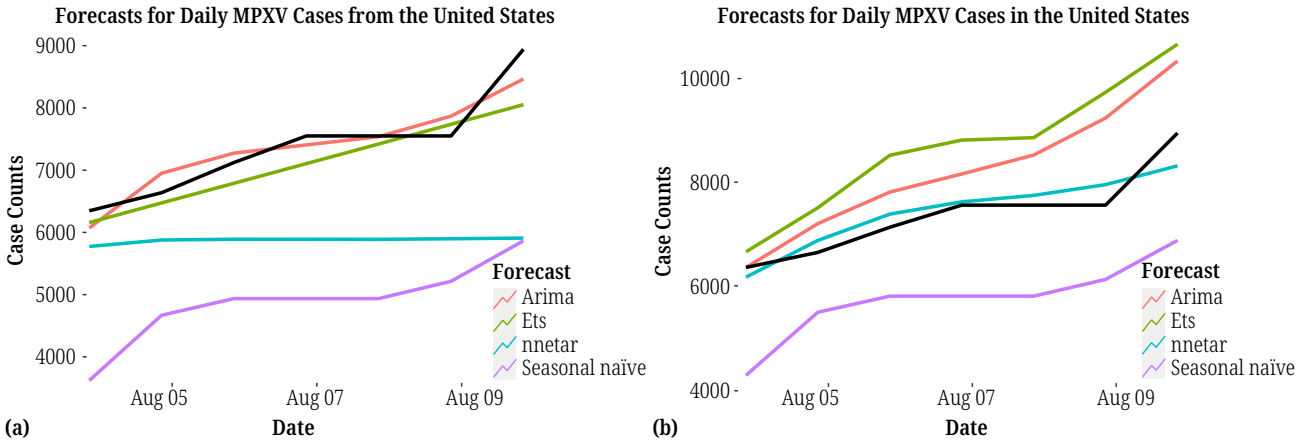

Fig. 11. The USA seven-day forecast from August 4, 2022, to August 10, 2022: (a) without Box-Cox transformation and (b) with Box-Cox transformation

Figures 7 to 11, the trajectory of MPXV cumulative cases for the five most-impacted countries is depicted. The figures on the left are untransformed, and the ones on the right are from the Box-Cox transformation. In Figure 7, France's MPXV cumulative cases plateaued from August 4, 2022, to August 8, 2022, and the trend in MPXV cumulative cases exceeded 25,000. The ARIMA and ETS models capture forecast distributions with higher accuracy. Figure 8 illustrates Germany's MPXV cumulative cases and it shows how it plateaued from August 5, 2022, to August 7, 2022, and the trend in MPXV cumulative cases exceeded 3000. The ETS model captured the forecast distribution with higher accuracy. Figure 9 reveals how Spain's MPXV cumulative cases plateaued from August 5, 2022, to August 8, 2022, and from August 9, 2022, with cumulative MPXV cases exceeding 5000. The NNETAR model captured the forecast distribution with higher accuracy. Figure 10 shows that the MPXV cumulative cases for UK plateaued from August 4, 2022, to August 8, 2022, and from August 9, 2022, with MPXV cumulative cases exceeding 3000. The NNETAR model also captures the forecast distribution with higher accuracy. In Figure 11, a steep trend is observed in the USA MPXV cumulative case test data with a brief plateau from August 5, 2022, to August 7, 2022. There was a consistent rise afterwards, and the cumulative MPXV cases exceeded 10,000. The NNETAR model in the USA forecasts trailed early from August 4, 2022, and drifted off-course after August 7, 2022.

This study shows how to optimize performance using the NNETAR, ARIMA, and ETS models with the Box-Cox transformation. The most noticeable difference in our investigation from other studies is that the MPXV cumulative case test data was implemented as a time-series for seven-day forecasts for the five most-affected countries as of August 7, 2022. Generalized time-series models are limited by their underlying assumptions, such as stationarity or the number of observations in the time series. An extensive test of hybridized A.R. and ANN models is widely accepted and justifiable for overall performance. The NNETAR model demonstrated comparable performance to the MPXV cumulative case test data when compared to ARIMA and ETS for the MPXV cumulative performance. In France, Germany, Spain, and the UK, persistence, or a long plateau, has revealed a decline in daily MPXV cases. This trend indicates that the response strategy to effectively contain the spread of MPXV cases in all four European countries. When compared to the ARIMA and ETS models, the NNETAR model with Box-Cox transformation exhibited higher accuracy for the USA MPXV test data. The trend of cumulative MPXV cases in the USA demands stringent measures to curb the increase in cases. Comparing our results with [9] and [10], we observed that gathering surveillance data is essential for tracking the cumulative MPXV case counts. Surveillance data plays a crucial role for understanding the dynamics of MPXV and reducing infections across regions. The crowdsourced ensemble probabilistic approach [10] for MPXV prediction serves as a valuable early tool for assessing the spread trajectories. One limitation of the approach adopted by [9] and [10] is the need for underlying compositions of the crucial features that are fundamental to time-series decomposition models.

4 CONCLUSION

A time-series analysis of the dynamic trajectory of MPXV outbreaks was conducted in the five most affected countries as of August 10, 2022. The machine learning models were trained using MPXV surveillance datasets from France, Germany, Spain, the UK, and the USA. We forecasted a seven-day horizon using NNETAR, ARIMA, ETS, and SNAIVE models. The hybrid NNETAR model, when applied with the Box-Cox

transformation, performed well in three countries. The models predicted that MPXV cases would plateau or decelerate in four European countries. In contrast, the ARIMA and ETS forecasts showed a steady increase in cumulative MPXV cases in the USA. We conclude that hybrid models with neural networks play a crucial role in time series forecasting. As a result, the investigation offers valuable insights to professionals and scientists who focus on epidemiological time series analysis, offering strategies for controlling the onset of an outbreak of MPXV.

The MPXV surveillance dataset comprises fewer than 100 observations from the five most-affected countries. For future experiments, a large number of MPXV observation datasets will be required to employ robust techniques and enhance forecast performance. MPXV can be analyzed using other machine learning ensemble types, including both linear and non-linear approaches. Furthermore, multivariate imputation techniques can be used to enhance the feature enrichment and improve the performance of MPXV predictions.

5 REFERENCES

[1] Y. Wu, Y. Yang, H. Nishiura, and M. Saitoh, "Deep learning for epidemiological predictions," *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, pp. 1085–1088, 2018. https://doi.org/10.1145/3209978.3210077

[2] P. Zarb et al., "The European centre for disease prevention and control (ECDC) pilot point prevalence survey of healthcare-associated infections and antimicrobial use," *Eurosurveillance*, vol. 17, no. 46, pp. 1–16, 2012. https://doi.org/10.2807/ese.17.46.20316-en

[3] P. Cawood and T. L. van Zyl, "Feature-Weighted Stacking for Nonseasonal Time Series Forecasts: A Case Study of the COVID-19 Epidemic Curves," *8th International Conference on Soft Computing and Machine Intelligence, ISCMI*, pp. 53–59, 2021. https://doi.org/10.1109/ISCMI53840.2021.9654809

[4] M. Ture and I. Kurt, "Comparison of four different time series methods to forecast hepatitis A virus infection," *Expert Syst Appl*, vol. 31, no. 1, pp. 41–46, 2006. https://doi.org/10.1016/j.eswa.2005.09.002

[5] X. Zhang, T. Zhang, A. A. Young, and X. Li, "Applications and comparisons of four time series models in epidemiological surveillance data," *PLoS One*, vol. 9, no. 2, pp. 1–16, 2014. https://doi.org/10.1371/journal.pone.0088075

[6] Q. Zeng et al., "Time series analysis of temporal trends in the pertussis incidence in Mainland China from 2005 to 2016," *Sci Rep*, vol. 6, pp. 1–8, 2016. https://doi.org/10.1038/srep32367

[7] B. Khayyat, F. Harrou, and Y. Sun, "Predicting COVID-19 Spread using Simple Time-Series Statistical Models," In *8th International Conference on ICT for Smart Society: Digital Twin for Smart Society, ICISS 2021 – Proceeding*, IEEE, 2021. https://doi.org/10.1109/ICISS53185.2021.9533192

[8] A. W. Rimoin et al., "Major increase in human monkeypox incidence 30 years after smallpox vaccination campaigns cease in the Democratic Republic of Congo," *Proc Natl Acad Sci U S A*, vol. 107, no. 37, pp. 16262–16267, 2010. https://doi.org/10.1073/pnas.1005769107

[9] T. McAndrew et al., "Early human judgment forecasts of human monkeypox, May 2022," *Lancet Digit Health*, vol. 4, no. 8, pp. e569–e571, 2022. https://doi.org/10.1016/S2589-7500(22)00127-3

[10] M. C. Martín-Delgado et al., "Monkeypox in humans: a new outbreak," *Rev Esp Chemoter*, vol. 35, no. 6, pp. 509–518, 2022. https://doi.org/10.37201/req/059.2022

[11] M. Kozlov, "Monkeypox in Africa: the science the world ignored," *Nature*, vol. 607, no. 7917, pp. 17–18, 2022. https://doi.org/10.1038/d41586-022-01686-z
[12] C. W. Langkop et al., "Multistate outbreak of monkeypox-Illinois, Indiana, Kansas, Missouri, Ohio, and Wisconsin, 2003," *Morbidity and mortality weekly report*, vol. 52, no. 27, pp. 642–546, 2003. https://pubmed.ncbi.nlm.nih.gov/12855947/
[13] J. B. Nuzzo, L. L. Borio, and L. O. Gostin, "The WHO declaration of monkeypox as a global public health emergency," *JAMA*, vol. 328, no. 7, pp. 615–617, 2022. https://doi.org/10.1001/jama.2022.12513
[14] P. Yuan et al., "Modelling vaccination and control strategies of monkeypox outbreaks at gatherings," *medRxiv*, 2022. https://doi.org/10.1101/2022.08.12.22278724
[15] G. E. P. Box and G. M. Jenkins, "Time Series Analysis: Forecasting and Control," Revised ed. San Francisco, CA, USA, 1976.
[16] Peter R. Winters, "Forecasting sales by exponentially weighted moving averages," *Management Science*, vol. 6, no. 3, pp. 231–362, 1960. https://doi.org/10.1287/mnsc.6.3.324
[17] A. Bhattacharyya, S. Chattopadhyay, M. Pattnaik, and T. Chakraborty, "Theta Autoregressive Neural Network: A Hybrid Time Series Model for Pandemic Forecasting," In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, IEEE, Shenzhen, China, pp. 1–8, 2021. https://doi.org/10.1109/IJCNN52387.2021.9533747
[18] M. Ngungu, E. Addai, A. Adeniji, U. M. Adam, and K. Oshinubi, "Mathematical epidemiological modelling and analysis of monkeypox dynamism with non-pharmaceutical intervention using real data from United Kingdom," *Front. Public Health*, vol. 11, 2023. https://doi.org/10.3389/fpubh.2023.1101436
[19] I. Priyadarshini, P. Mohanty, R. Kumar, and D. Taniar, "Monkeypox outbreak analysis: An extensive study using machine learning models and time series analysis," *Computers*, vol. 12, no. 2, p. 36, 2023. https://doi.org/10.3390/computers12020036
[20] A. Bleichrodt, S. Dahal, K. Maloney, L. Casanova, R. Luo, and G. Chowell, "Real-time forecasting the trajectory of monkeypox outbreaks at the national and global levels, July–October 2022," *BMC Med*, vol. 21, no. 1, p. 19, 2023. https://doi.org/10.1186/s12916-022-02725-2
[21] W. Wei et al., "Time series prediction for the epidemic trends of monkeypox using the ARIMA, exponential smoothing, G.M. (1, 1) and LSTM deep learning methods," *Journal of General Virology*, vol. 104, no. 4, 2023. https://doi.org/10.1099/jgv.0.001839
[22] B. Long, F. Tan, and M. Newman, "Forecasting the monkeypox outbreak using ARIMA, prophet, NeuralProphet, and LSTM models in the United States," *Forecasting*, vol. 5, no. 1, pp. 127–137, 2023. https://doi.org/10.3390/forecast5010005
[23] CDC, "U.S. Monkeypox Case Trends Reported to CDC," 2022. https://www.cdc.gov/poxvirus/monkeypox/response/2022/mpx-trends.html (accessed August 2, 2022).
[24] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2. pp. 211–243, 1964. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x
[25] V. M. Guerrero, "Time-series analysis supported by power transformations," *J Forecast*, vol. 12, no. 1, pp. 37–48, 1993. https://doi.org/10.1002/for.3980120104
[26] R. J. Hyndman and G. Athanasopoulos, "Time Series Cross Validation," (3rd ed), *Otexts: Melbourne, Australia*, 2021. https://otexts.com/fpp3/tscv.html (accessed May 25, 2022).
[27] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," *Geosci Model Dev*, vol. 7, no. 3, pp. 1247–1250, 2014. https://doi.org/10.5194/gmd-7-1247-2014
[28] M. M. Dancho, D. Vaughan, and M. M. Dancho, "Package 'time to,'" 2022.

6 AUTHORS

Solomon Oluwole Akinola received a Bachelor of Computer Engineering degree from the Ladoke Akintola University of Technology. He received his master's degree from the University of Ibadan. He is currently pursuing his PhD at the University of Johannesburg. His research interests include applying machine (deep) learning to natural language processing, intelligent user interfaces, and the Internet of Things. (E-mail: oluwolea@uj.ac.za).

Qing-Guo Wang received a B.Eng. in Chemical Engineering in 1982, M. Eng. in 1984 and a PhD in 1987, both in Industrial Automation, all from Zhejiang University, PR China. He held the Alexander-von-Humboldt Research Fellowship in Germany from 1990 to 1992. From 1992 to 2015, he was with the Department of Electrical and Computer Engineering at the National University of Singapore, where he became a full-time professor in 2004. He is a Distinguished Professor at the Institute for Intelligent Systems, University of Johannesburg, South Africa. He holds an A rating from the National Research Foundation of South Africa (NRF). He is a member of the Academy of Sciences in South Africa. His current research interests are mainly in modelling, estimating, predicting, controlling, optimising, and automating complex systems, including but not limited to industrial and environmental processes, new energy devices, defence systems, medical engineering, and financial markets. He has published nearly 300 international journal papers and 7 research monographs. He received approximately 15000 citations with an h-index of 65. He is currently the Deputy Editor-in-Chief of the ISA Transactions (USA). (E-mail: wangq@uj.ac.za).

Peter Olukanmi obtained a PhD from the University of Johannesburg and an MSc in Computer Science from the University of KwaZulu-Natal (UKZN) BSc in Systems Engineering from the University of Lagos. He won two IEEE conference awards in soft computing and machine intelligence. His research interests include fundamental and applied data science and mathematical modelling. (E-mail: polukanmi@uj.ac.za).

Tshilidzi Marwala is a South African mechanical engineer and a computer scientist. He became a Professor at the University of the Witwatersrand in 2003 and a chairperson of System and Control Engineering in South Africa. He had previously worked at the CSIR and South African Breweries. Marwala's research interests include the theory and application of artificial intelligence in engineering, computer science, finance, economics, social science, and medicine. Marwala has made fundamental contributions to engineering science, including developing the concept of pseudo-modal energies and proposing the theory of rational counterfactual thinking, reasonable opportunity cost, and flexibly bounded rationality. (E-mail: tmarwala@uj.ac.za).