Health and Technology
https://doi.org/10.1007/s12553-023-00797-5

ORIGINAL PAPER

COVID-19 vaccine prediction based on an interpretable CNN-LSTM model with three-stage feature engineering

Lubna Altarawneh[1] · **Hao Wang**[1] · **Yu Jin**[1]

Received: 25 August 2023 / Accepted: 22 October 2023

Abstract

Purpose Vaccine supply planning remains a critical and challenging issue for developing countries due to the limited resources and fluctuating vaccine demand, reflected by the sudden increase and decrease in the number of vaccinated people. Therefore, it is essential to accurately predict vaccination rates and interpret the prediction results to anticipate vaccine demand. Previous research focused on the prediction of vaccination rates by considering vaccination data and overlooked the importance of including features that highly affect the prediction results.

Methods This research proposes an interpretable Convolutional Neural Network and Long Short-term Memory (CNN-LSTM) model for COVID-19 vaccination rate prediction, specifically for developing countries. The proposed model includes a three-stage feature engineering process that utilizes domain knowledge to construct new features of disease spread and the strictness of government policies. The temporal correlation of the features is captured to improve the prediction accuracy.

Results The case studies were conducted on Nigeria, Nepal, and Peru's datasets to ensure the model's effectiveness and generalizability. The results indicate the effectiveness of the model, showing an average 30% decrease in the Mean Absolute Percentage Error (MAPE) with the enhanced feature engineering process and 16% additional improvement after combining the CNN with LSTM compared to the standalone LSTM model.

Conclusion The interpretability of the CNN-LSTM model is a key aspect considered in this research to explain the prediction results. The model's interpretability aims to identify important features affecting the prediction results and provide government and vaccine companies with guidelines in the decision-making for vaccine supply planning in developing countries.

Keywords Vaccination rates · Vaccine demand · Hybrid CNN-LSTM · Interpretable model

1 Introduction

Respiratory infectious diseases pose a continuous threat to society. Being contagious in nature, their transmission results in a large number of infections and high mortality rates [1]. Different respiratory infectious diseases emerged throughout the years, including different variants of influenza. Most recently, the emergence of COVID-19 raised global concerns and had significant social and economic impacts with different variants of the virus being detected. Therefore, interventions must be implemented to reduce the transmission rates of such diseases.

Different pharmaceutical and non-pharmaceutical interventions are implemented; however, vaccination remains the most effective intervention [2]. At the beginning of an outbreak, enormous efforts are taken to manufacture, supply, and distribute vaccines. However, with the limited supply at the beginning of a crisis, vaccines cannot be distributed globally at the same time. Therefore, it is vital to accurately predict vaccine demand to ensure an adequate vaccine supply. Specifically, with the high demand that can rise for vaccines, enough vaccines should be produced to meet global or regional demand. On the other hand, oversupply should be avoided due to the high inventory cost and the vaccines being considered as perishable items.

This article is part of the *COVID-19 Health Technology: Design, Regulation, Management, Assessment*

Yu Jin
yjin@binghamton.edu

Lubna Altarawneh
ltarawn1@binghamton.edu

Hao Wang
hwang147@binghamton.edu

[1] Department of Systems Science and Industrial Engineering, Binghamton University, SUNY 4400 Vestal Parkway East, Binghamton, NY 13902, USA

Previous research considered historical vaccination data to predict COVID-19 vaccine uptake using Auto-regressive Integrated Moving Average (ARIMA) [3, 4]. Factors related to the number of cases, infectious disease spread, and the growth and recovery of the cases can impact vaccination rates based on recent research. In addition, government policies and procedures can have a direct effect on vaccination rates. Such factors, which have not been considered in previous research, can drastically affect the prediction results. Therefore, it is important to incorporate domain knowledge and conduct appropriate feature engineering to determine indispensable factors that can substantially improve prediction accuracy and help operational planning of vaccine supply. In this study, a three-stage feature engineering process is implemented to not only extract features related to disease positive cases and deaths, but also expand features to include infectious disease spread and strictness of government policies that affect the vaccination rates for a region or a country.

Moreover, statistical forecasting methods such as ARIMA may produce reliable results for short-term prediction, but it is challenging to model complex nonlinear and nonstationary relationships between variables when long-term prediction is needed. Hence, they are not able to completely capture and learn the behavior of the data [5]. Recent advancements in machine learning and deep learning, such as Long Short-Term Memory (LSTM) networks, have shown their capability in capturing such complex temporal auto-correlations and dependency of time series data and outperform statistical models including ARIMA [6–8]. Since LSTM was originally proposed in [9] to overcome the gradient vanishing problem faced by traditional Recurrent Neural Networks (RNN), it has been used in applications such as flight trajectory prediction [10], Taxi demand [11, 12], ride-hailing service demand [13], and traffic prediction [14].

However, LSTM suffered from limitations when dealing with multivariate data where the multivariate correlations are not fully considered with the original structure. Several works improved the performance of LSTM for multivariate prediction by incorporating Convolutional Neural Networks (CNN) to account for the correlation between multivariate variables and extract important information to improve the features constructed for the prediction.

In the case of vaccine demand prediction, it is desired to acquire long-term prediction results to ensure vaccine production and supply planning in a timely manner. Therefore, in this study, a hybrid CNN-LSTM model is proposed to predict vaccination rates on a weekly, bi-weekly, and monthly basis and provide vaccine demand for a given realm. Its performance is compared to ARIMA, the LSTM purely based on the historical vaccination rate, and the LSTM based on both historical vaccination rate and selected features to show how the extracted and convoluted features can improve the prediction results.

Despite the outstanding results machine learning models produce, most models lack transparency and interpretability about their results because of their black-box nature [15]. Therefore, this research focuses on the interpretability of the proposed CNN-LSTM model to assist government and vaccine companies with decision-making by applying an Explainable Artificial Intelligence (XAI) system to overcome the black-box nature of machine learning models. Shapley Additive Explanations (SHAP) methodology is used in XAI to derive disease growth patterns and government strictness policies that affect the vaccination rates for a country.

The research objectives can be summarized as follows: 1) Identify the different factors affecting vaccination rate for a country or region including the temporal factors related to COVID-19 and the spatial factors related to a country's regulations in terms of COVID-19; 2) predict the vaccination rates for a country or region using its historical vaccination data and by incorporating the related factors data; 3) explain the prediction results and the factors contributing to the prediction results; 4) anticipate vaccine demand by identifying expected vaccination rates for a country or region; and 5) help vaccine manufacturers with planning vaccine supply and distribution based on the anticipated demand.

Accurate and explainable forecasting of vaccination rates, especially for developing countries can ensure on-demand vaccination requests. For such countries with limited resources, vaccine supply and distribution raise increased challenges due to the high population growth. In addition, it experiences a lack of proper storage with appropriate temperature for vaccines. Therefore, it is important to provide interpretable and accurate predictions of vaccination rates and identify vaccine demand in developing countries so that vaccines are supplied based on the anticipated demand. The contributions of this paper are summarized as follows:

- A three-stage feature engineering process, which consists of feature set extraction and construction, features preprocessing, and feature selection, is conducted to utilize domain knowledge and identify factors that affect vaccination rates in a country or region.
- A hybrid CNN-LSTM model, which takes the advantages of CNN and LSTM in extracting critical information and capturing complex temporal dependence of multivariate data, is developed to predict vaccination rates based on the selected features.
- An XAI system is designed, which is inherently explainable and can understand the predictions of the black-box CNN-LSTM model and use it for the decision-making process. The interpretability of the proposed model is considered both locally for a single data point and globally over the entire dataset to assist policymakers and

vaccine manufacturers in understanding the related factors contributing to the prediction results.

To the best of our knowledge, this is the first attempt to identify factors affecting COVID-19 vaccination rates for developing countries and use it to enhance the prediction as well as interpret the different factors effects on the prediction results to assist policy makers in their decision making process. The remainder of this paper is organized as follows. Section 2 describes a summary of recent related studies. Section 3 describes the methodology used in this research. Section 4 provides experimental results. Section 5 concludes with the findings and discusses future work.

2 Literature review

There has been an increasing demand for vaccines due to the spread of infectious diseases and outbreaks in recent years. This signifies the importance of anticipating vaccine demand in a timely manner. Many studies used conventional statistical models to predict vaccination rates and vaccine demand. For instance, ARIMA was used to predict vaccine demand at different health centers based on vaccine stock history to avoid shortage and excess stock [16]. ARIMA was also used to predict vaccine demand based on historical data for a population receiving the vaccine [17]. Most recently, ARIMA was used to predict COVID-19 vaccine uptake [3]. Tamilselvan et al. [4] suggested vaccination distribution methods to end lockdown measures and predicted vaccination rates using Prophet. The experimental results indicate an increase in vaccination rates in the future indicating the need for lockdown measures. Arifin and Anas [18] forecasted the COVID-19 vaccination rates for Indonesia by considering past data and by using Prophet and simulation to determine vaccine plans. The experimental results were compared to the actual results to derive recommendations to improve vaccination rates in Indonesia. Zheng et al. [19] forecasted the daily demand for COVID-19 vaccines based on the number of appointments booked and infection rates using a fuzzy machine learning model, the forecasted demand was used to improve vaccine distribution using an evolutionary algorithm. Rathinasamy and Kandhasamy [20] predicted vaccination rates on a country-to-country basis using ARIMA. The experimental results show the state of COVID-19 vaccine administration and estimate vaccination rates for different countries. Different measures were used to assess the performance of the ARIMA prediction model including Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Other studies used Mean Absolute Percentage Error (MAPE) to provide a scale-independent assessment of the ARIMA model [21, 22]. A summary of recent studies on COVID-19 prediction is shown in Table 1.

Although statistical time-series forecasting models can provide predictions of vaccine demand and uptake, they often consider univariate historical time series only for making predictions and do not take into consideration different features that might affect the performance of the prediction model. In addition, they have limited capability in capturing the complex nonlinear and nonstationary temporal dependency of time series. Several studies identified the correlation between COVID-19 vaccination and the number of cases and deaths [26, 27]. Therefore, COVID-19 confirmed cases and deaths data [28] are considered as potential candidate features in this research. In addition, we further consider testing features in the COVID-19 testing data [29], as research proved testing and vaccination to be correlated as well [30]. The relationships between these features and vaccination rates are explored to select features that have a high impact on vaccination rates.

Multivariate time series prediction has been increasingly used in different applications as it results in improved prediction results compared to univariate time series prediction. It has been used for COVID-19-related forecasting where several features were considered to improve the prediction results. Different features related to lockdown measures have been used to predict COVID-19 cases for a number of countries having similar demographic and socioeconomic aspects [31]. The number of COVID-19 hospital Intensive Care Units (ICU) in Brazil was also forecasted by considering features such as the current number of daily cases, vaccination rates, non-pharmaceutical interventions, social isolation index, and regional hospital beds occupation [32]. Similar to the previous studies of COVID-19, the prediction of COVID-19 vaccine rate should also incorporate different features that substantially affect the vaccine uptake or hesitancy. To the best of our knowledge, none of the previous studies provided the prediction results based on the important features.

Before including the features in the prediction, it is essential to evaluate and decide which features should be selected. Previous research which predicted the number of COVID-19 cases and hospital ICU used correlation analysis to identify related features [31, 32]. However, correlation only reflects one aspect of the importance. Constructing and conducting an efficient feature engineering process enables the use of critical features to improve predictions and is often overlooked in many applications.

In the case of COVID-19 vaccination, including a feature engineering process is essential to predict vaccination rates in order to identify factors resulting in increased vaccine uptake or vaccine hesitancy, and assist policymakers.

Feature selection methods have been extensively used to determine the best subset of features to use for prediction problems. Identifying an efficient subset of features helps improve the prediction results and reduces

Table 1 COVID-19 prediction related studies

Study	Reference	Purpose	Method
Number of cases	Ribeiro et al. (2020) [23]	Predict cumulative COVID-19 cases	ARIMA, CUBIST, RIDGE, RF, and SVR
	Tamang et al. (2020) [24]	Predict number of COVID-19 cases and deaths	Artificial neural network-based curve fitting techniques
	Said et al. (2021) [31]	Predict daily COVID-19 cases by considering lockdown measures	Bidirectional LSTM
	Somyanonthanakul et al. (2022) [25]	Predict number of COVID-19 cases for efficient crisis management	ARIMA and Association Rule Mining (ARM)
Number of hospital ICU	Borges and Nascimento (2022) [32]	Predict the ICU demand by considering daily cases, vaccination rates, and other factors	Prophet-LSTM
Vaccine demand and distribution	Zheng et al. (2022) [19]	Forecast vaccine demand and optimize vaccine distribution	Fuzzy machine learning model and Evolutionary algorithm
Vaccination rate	Tamilselvan et al. (2021) [4]	Suggest vaccination distribution methods to end lockdown measured by predicting vaccination rates	Prophet
	Arifin and Anas (2021) [18]	Develop a vaccination plan based on forecasted vaccination rates	Simulation and Prophet
	Rathinasamy and Kandhasamy (2022) [20]	Predict vaccination rates for countries	ARIMA
	Zhou and Li (2022) [3]	Predict vaccine uptake based on clinical data and web search queries	ARIMA

the computational time for training the prediction model. There are two well-known methods for feature selection which are filter methods and wrapper methods. Filter methods use the inherent properties of the features by using statistical means to measure them, while wrapper methods use the model on different subsets of features and compare the model's predictive performance to identify the best set of features [33]. Two filter methods of mutual information and relief-f have been used to select features to predict COVID-19 using chest CT images [34]. The wrapper method of Recursive Feature Elimination (RFE) has been used to select a subset of features to use for COVID-19 prediction [35]. Although wrapper methods provide better results in terms of the selected features, they are computationally costly. Therefore, in the case where there is a large number of features, filter methods can be used first for preliminary screening as they are computationally efficient followed by wrapper methods. This takes advantage of the computational efficiency of filter methods and the competitive results the wrapper methods produce in terms of the selected features. A combination of filter and wrapper methods for feature selection has been used in the literature to improve the feature selection process. Filter and wrapper methods were used to predict the COVID-19 growth rate in India [36]. They were also used to predict mortality rates for patients with COVID-19 [37].

Feature selection methods are commonly used when a list of features is available. However, when there is large input data, some features need to be constructed by combining or transforming certain features to reduce the feature set size. Feature construction involves transforming input data by reducing its size. It constructs more representative features without distorting the original relationship between features. It is useful when there is a large number of features that needs to be reduced without losing important information. Feature construction has been used to construct a new feature that combines wind strength and direction components in a single feature to be used for wind power forecasting [38]. It was also used to create time features to predict wind power assuming that wind power is associated with time features such as the day of the week or hour of the day [39]. Feature construction has been extensively used in the healthcare domain due to the large amount of available data [40]. Features are often constructed by utilizing domain knowledge from experts [41]. This work utilizes domain knowledge to extract and include important information related to vaccination rates in multivariate time series prediction of vaccination rates to include critical information not present in the original dataset.

The features selected are used as input to the prediction model to improve its results. Statistical forecasting models commonly include historical data as a feature for prediction. In order to incorporate multiple features in the time series prediction, recent studies proposed using machine learning and deep learning models to overcome the limitations of the conventional statistical time series forecasting models. For instance, a Recurrent Neural Network (RNN) for modeling multivariate spatial temporal sequences was proposed in [42]. A special type of RNN, LSTM, which has a special structure to handle long-term dependencies in the data, is applied, and is in the center of research as it has been extensively used for multivariate prediction. Multivariate LSTM has been used to predict the number of new and total COVID-19 cases and outperformed univariate LSTM for one-day ahead prediction [43]. Multivariate LSTM with environmental parameters outperformed univariate LSTM in predicting daily cases of COVID-19 for several cities [44].

Several works integrated CNN with LSTM to extract important information from features to be included in the LSTM prediction model. For instance, multivariate CNN-LSTM was used to predict mobility in the COVID-19 pandemic and outperformed both standalone CNN and LSTM in obtaining the prediction results [45]. Hybrid CNN-LSTM provided a significant boost in forecasting the number of confirmed COVID-19 cases compared to the standalone models [46]. It has not be used to predict vaccination rates or vaccine demand for COVID-19. In this work, CNN-LSTM is used for long-term vaccination rate prediction of COVID-19. It is critical for the vaccine supply chain operation to estimate the vaccine demand level in advance to allow timely planning and distribution of vaccines. Therefore, vaccination rates are predicted on a weekly, bi-weekly, and monthly basis to allow vaccine manufacturing and distribution planning in a timely manner.

The main contribution of this work is to develop a CNN-LSTM model to predict vaccination rates for a country or region that is inherently interpretable and can explain the different features affecting the prediction results. Explainable AI has been used to explain the prediction results for LSTM in several applications to assist with decision-making [47], but has not been considered in explaining CNN-LSTM model performance, especially for vaccination rates prediction. Explainable AI is used in applications to assist policy makers in understanding how the model reached its predictions and helps with decision-making and planning [48]. It answers questions related to model transparency which are of greater consequences in the healthcare area in which decisions are directly affecting people's lives [49]. Specifically, for vaccination rates, it is essential to understand the factors resulting in vaccination uptake and allow actions to be derived from understanding the model's prediction decisions.

3 Methodology

The main goal of this study is to accurately predict vaccination rates for a country or region so that it can be used to identify vaccine demand and assist policy makers including government and vaccine manufacturers. This is achieved by developing an interpretable prediction model for vaccination rates with an enhanced feature engineering process that enables the use of relevant features to provide accurate predictions and better anticipate vaccine demand. In order to achieve this goal, a three-stage feature engineering process that utilizes domain knowledge for feature construction, feature preprocessing, and feature selection by filter and wrapper is developed. Then a hybrid CNN-LSTM prediction model is developed where the selected features are used as input to obtain long-term predictions. The following sections provide a detailed discussion of the methodology used to obtain the prediction results for vaccination rates and identifying vaccine demand for countries and regions.

3.1 A three-stage feature engineering process

The prediction of vaccination rates is considered as a spatial temporal forecasting problem where both spatial and temporal characteristics of the series are captured. Spatial temporal forecasting considers both temporal information based on historical records and spatial variability according to the specified locations when predicting future values [50]. In the case of vaccination rates, temporal modeling considers historical vaccination rates and external temporal features, where spatial modeling considers country-specific regulations and procedures related to COVID-19 that can potentially affect the country's vaccination rates. In order to integrate critical features in the prediction, a three-stage feature engineering process is developed as illustrated in Fig. 1.

The proposed feature engineering process consists of three main steps. First, a feature set is established to include 1) the extracted disease positive cases, deaths, and testing, and 2) the new features constructed to represent disease spread and strictness of government policies using domain knowledge. Second, the extracted and constructed features, as well as the historical vaccine rates, need to be preprocessed by cleaning the null or redundant values and normalization to remove the scale effects. Third, the best subset of features is selected by applying filter methods for preliminary screening and wrapper methods to select optimum features out of the candidate features. The three stage feature engineering process proposed to improve the vaccination rate prediction enhances on the typical feature selection method for vaccination rate prediction in several aspects. First, it incorporates constructed features that contain new information related to COVID-19 not present in the original

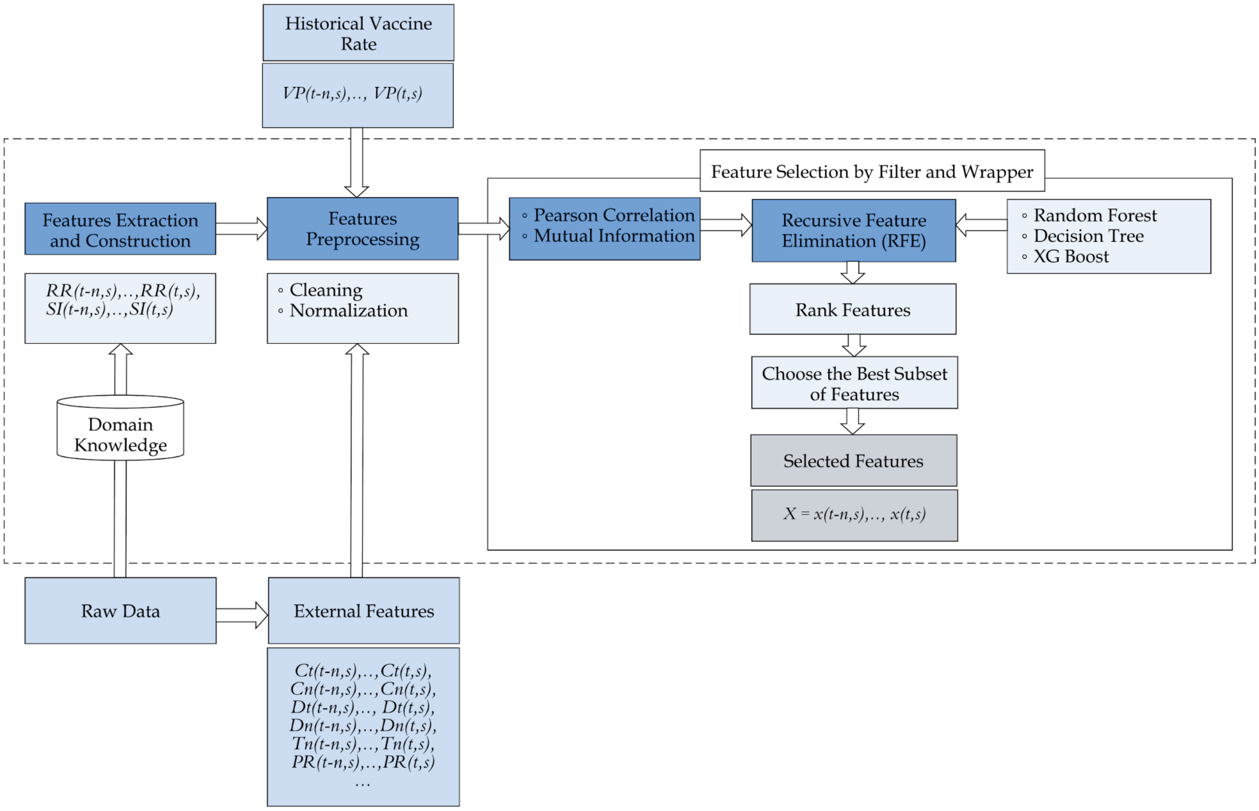


Fig. 1 Three-stage feature engineering process

features such as reproduction rate and stringency index. The features constructed are considered as candidate features in the feature selection process. The reproduction rate adds more information related to how the disease spreads in the population compared to only considering the number of confirmed cases which does not include information about the disease transmission. Also, the stringency index incorporates information related to the government response to COVID-19 in terms of school, workplace, and public events closures as well as other indicators, which adds important information to how the government is reacting to the COVID-19 situation. The domain knowledge of the disease and government behavior utilized to construct these features can enhance the vaccination rate prediction since new and important factors not considered in typical feature extraction is found. Second, the features constructed and extracted undergo rigorous preprocessing where a feature with null values for the entire period taken is removed while a feature with null values for specific days is filled by considering the average values for the feature for the remaining days. There are two cases observed with missing feature values data, at the beginning and at the end of the time period taken. For the first case, the observations in the near history are considered as reference to fill the missing data. For the second case, the observations in the near future are considered as reference to fill the missing data. This ensures that the estimated feature values are not all based on average but also based on the situation of COVID-19 and feature values around each period. Third, the feature selection used in this process includes two phases for accurate identification of the optimum features. It uses two filter methods for preliminary screening of the features and to reduce the feature set in the first phase, where a wrapper method is performed in the second phase that selects the optimum features out of the candidate features by using three different methods for accurate identification of the features. The combination of the feature selection phases of filter and wrapper methods while incorporating several methods at each phase aims to identify the optimum features affecting vaccination rate when considering the constructed and extracted features. By incorporating the selected features into the prediction, the spatial and temporal dependency can still be captured whenever the region and time change. It is worth noting that the continuous expansion of feature sets cannot always provide a performance boost. Instead, it may result in performance degradation due to the overfitting for the modeling training [51]. Therefore, effective feature engineering and selection are needed to acquire the optimum features.

3.1.1 Features extraction and construction

Considering the objective of predicting the vaccine rate and identifying vaccine demand, the vaccination rates VP collected from country s are set as the target feature. The corresponding historical record of the target feature VP from time $t - n$ to the current time t is included in the feature set, where n denotes the prediction period for which historical n days are used to predict future n days. Moreover, several research identified the correlation between COVID-19 vaccination and the number of cases, deaths, and tests as discussed in the literature. Therefore, the data related to COVID-19 confirmed cases, deaths, and testing of country s are considered potential candidate features and are collected from publicly available records [52]. The features considered are summarized in Tables 9, 10, and 11 in Appendix A. The relationships between these features and vaccination rate will be explored in Section 3.1.3 to select features that have a high impact on the target.

In addition to the features directly extracted from the COVID-19 public record, the reproduction rate $RR(t, s)$ of COVID-19 is constructed to include the domain knowledge of COVID-19 transmission, and changes when the population becomes immunized or vaccinated [53]. It identifies disease spread among the population at a given time and location by considering features related to the number of cases. World Health Organization (WHO) estimates COVID-19 reproduction rate using mathematical models [54], which although seems simple but is challenging to track. The estimation process of $RR(t, s)$ includes three main steps [55]. First, data related to new COVID-19 cases are used to construct a time series that represents the number of infected individuals $I(t, s)$ in a location or region over time. $I(t, s)$ is initialized by $I_0 = C_0$ where C_0 is the total number of infectious cases at the beginning of the time frame. Then subsequent values of $I(t, s)$ are constructed recursively based on the transition rate γ from infectious to recovered and the number of new cases $Cn(t, s)$ as shown in Eq. 1. A growth rate GR of $I(t, s)$ is estimated for this time series using a Kalman filter as shown in Eq. 2, where γ indicates the transition rate from infectious to recovered. Next, $RR(t, s)$ is obtained from the estimated GR in Eq. 3. The estimation is performed by the Kalman smoother [56], which provides optimal estimates of $RR(t, s)$ given the full sample information of GR. This is assuming that the data is generated using Susceptible Infectious Removed (SIR) model. The equations utilize domain knowledge about new cases to produce a new feature of $RR(t, s)$ which provides critical information to identify the growth of COVID-19 cases when the population becomes immunized or vaccinated.

$$I(t, s) = (1 - \gamma)I(t - 1, s) + Cn(t, s) \tag{1}$$

$$GR(I(t, s)) = \gamma(RR(t, s) - 1) + \epsilon_t, \epsilon_t \sim i.i.d.\mathcal{N}(0, \sigma_\epsilon^2) \tag{2}$$

$$RR(t, s) = RR(t - 1, s) + \eta_t, \eta_t \sim i.i.d.\mathcal{N}(0, \sigma_\eta^2) \tag{3}$$

Domain knowledge of existing government rules and restrictions related to COVID-19 is also utilized to construct Stringency Index $SI(t, s)$ which is calculated by government response tracker. Government restrictions and public campaigns have been widely affecting people's choice to take vaccines and are therefore considered in this research. $SI(t, s)$ indicates the strictness of government policies, which is a composite measure based on M Response Indicators $RI(t, s)$ shown in Fig. 2. The index on any given day is calculated as the mean score of the nine metrics, each taking a value between 0 and 100, with 100 being the strictest as shown in Eq. 4. The metrics mentioned could potentially affect the vaccination rates for a population in a given region, with government policies and public campaigns indirectly affecting an individuals' choice to take vaccines. The equations for both $RR(t, s)$ and $SI(t, s)$ are used to construct a new feature space with low dimension.

$$SI(t, s) = \frac{1}{M}\sum_{m=1}^{M} RI(t, s) \tag{4}$$

3.1.2 Features preprocessing

The features extracted and constructed need to be preprocessed before further treatment. Therefore, missing values are handled using different approaches depending on where the missing data is located. First, features with null values for a specific country are removed. Then, features with missing entries for a specific date are estimated by taking the average for that feature to ensure no null values are included. Features with null values at the beginning of the time period taken, such as $RR(t, s)$, are estimated by taking Next Observation Carried Backward (NOCB) since it would not be reasonable to take the average of data in this case. Features with null values at the end of the time period taken are estimated by taking Last Observation Carried Forward (LOCF) so it can represent the latest changes in COVID-19 situation. Redundant features are removed as they may contain similar information. Each remaining feature x is normalized based on Eq. 5 to ensure a similar scale and to keep all relationships between features with no bias present.

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{5}$$

3.1.3 Feature selection by filter and wrapper

After the first two steps of feature engineering are performed, it is expected that there are features that need to be

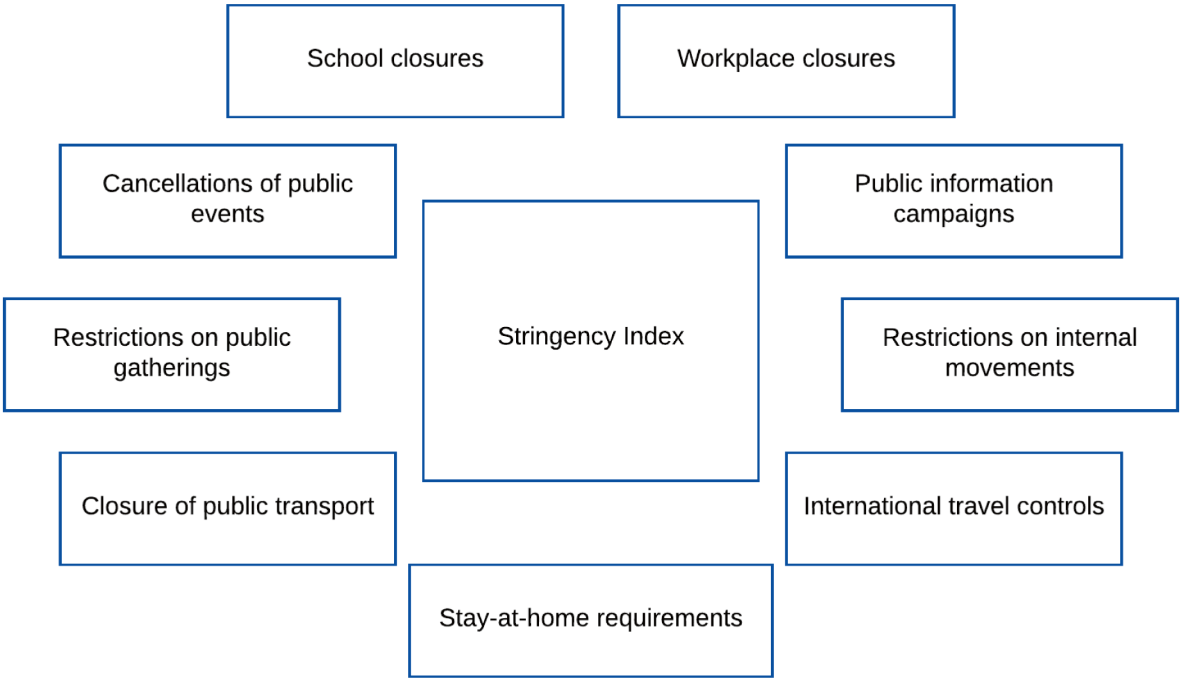


Fig. 2 Response Indicators used in calculating Stringency Index

further evaluated to avoid potential over-fitting problems. The proposed feature selection process of filter and wrapper methods is illustrated in Fig. 3. The effectiveness of using filter and wrapper methods to determine the best subset of features is discussed in previous research [57, 58]. First, filter methods are applied to assess the relevance of the features and then eliminate features with low ranks [59]. Filter methods are known to be fast, scalable, and independent of the model used. The filter in the proposed framework is set based on both Pearson correlation and mutual information to capture the feature dependency from different perspectives. Features with low correlation (< 0.3) with the target feature and low mutual information (< 0.5) are removed, and features with medium and high correlation or high mutual information are considered as candidates. The two filtering methods of Pearson correlation and mutual information are considered as the preliminary screening of the features to reduce the computational cost when applying the wrapper method. A wrapper method is then used to select the best subset of features from the candidate features. The steps of the wrapper method used, RFE, are illustrated in Algorithm 1. RFE selects the best feature subset based on the

```
Input a set of training samples with feature dimension $d$; a training algorithm
Initialize the list of candidate features $X_{in} \leftarrow \{1, . . ., d\}$
Initialize the list of eliminated features $X_{out} \leftarrow \emptyset$
Determine appropriate number of features $n$
While $X_{in} \neq \emptyset$ do
    Train the model with features in $X_{in}$
    Calculate the features ranking based on ranking criteria
    Sort $X_{in}$ according to the descending order of the ranking criteria
    Remove feature with least ranking
    $X_{in} \leftarrow X_{in}$ - feature with least ranking in $X_{in}$
    Stop when number of features in $X_{in} = n$
End while
Output a ranked list of features $X$
```

Algorithm 1 Recursive Feature Elimination (RFE)

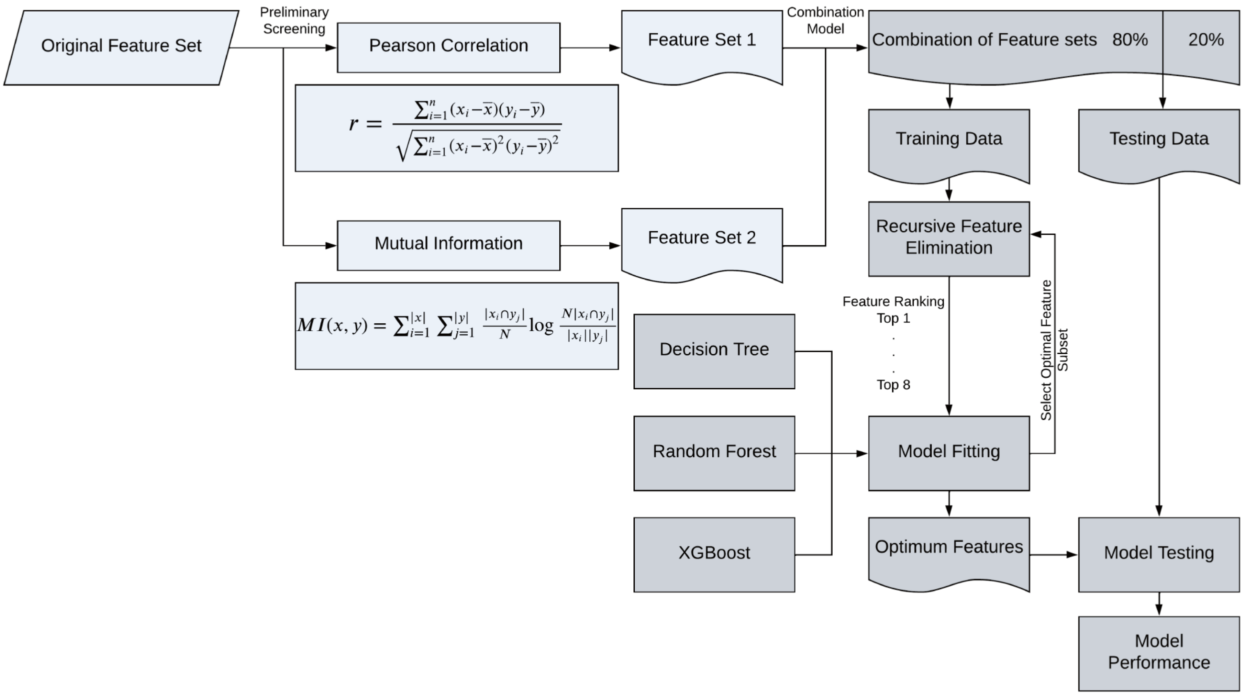


Fig. 3 Feature selection by filter and wrapper

model and prediction accuracy. It starts with all candidate features and sequentially removes features that reduce the prediction accuracy. A threshold should be set for the number of features to use. Three different models are considered when performing RFE, which are Random Forest, Decision Tree, and XGBoost, to verify that the resultant features have a significant impact on the target feature.

3.2 Proposed CNN-LSTM prediction model

In this work, CNN and LSTM are used for long-term prediction. The CNN provides preprocessing layers that have the ability to filter raw input data and extract important information to construct new feature space. Although it is commonly used to extract key information from images [60], but it has been used to construct new feature space from time series data in recent years [61]. The convolutional kernel slides over the input features constructed as matrix multiple times and applies convolution operations each time to form several convolved features that represent the feature space better than the original features. The Rectified Linear Unit (ReLU) is then applied to enable the model to learn complex mapping functions and help avoid the vanishing gradient problem, followed by a max pooling layer. The max pooling layer slides windows multiple times to extract maximum value from the convolved features every time the window slides to produce a lower dimension matrix. The pooling layer constructs a summary of the convolved features that the convolutional layer produced and helps reduce the features map [62].

Next, the features extracted and summarized using CNN convolutional layer and max pooling layer are fed into the LSTM layer through gate units. The gate units comprises of input, output, and forget gate. The input gate controls the contribution of new information to the cell, that is, the features smoothed by the convolutional layers. The output gate controls the weight of the information in the cell that is used for the output activation, which determines the features information to be kept. The forget gate controls the proportion of information that remains in the cell, which are the features that do not affect the vaccination rates prediction and do not need to be considered. A LSTM network computes the vaccination rates $VP = (VP(t+1,s), ..., VP(t+n,s))$ of country s of n days in the future based on an input sequence $X = (x(t-n,s), ..., x(t,s))$ that represents the selected features by iteratively calculating the network unit activations using Eqs. 6-11, where W is the weight matrices, b is the bias vectors, σ is the logistic sigmoid function, i is the input gate, f is the forget gate, o is the output gate, c is the cell activation vectors, m is the cell output activation, and $\odot$ is the element-wise product of the vectors. The LSTM is able to store important characteristics extracted from the CNN layers and preserves long-term memory by the use of memory cells [63]. Memory cells will update the states every time when a gate unit is activated. The hidden state of the LSTM keeps updating after a number of steps and utilizes feedback connections. Therefore, LSTM overcomes the challenges faced by Recurrent Neural Networks (RNN) of vanishing gradient problems. A flatten layer is added to reduce the dimensions, which is followed by a dense layer to obtain the prediction results. The architecture of the proposed CNN-LSTM model is illustrated in Fig. 4 with different CNN and LSTM layers. Several experiments and runs are conducted to identify the best parameter settings and configurations. A summary of the detailed model specifications and parameters is shown in Table 2.

$$i(t) = \sigma(W_{ix}x(t) + W_{im}m(t-1) + W_{ic}c(t-1) + b_i) \tag{6}$$

$$f(t) = \sigma(W_{fx}x(t) + W_{fm}m(t-1) + W_{cf}c(t-1) + b_f) \tag{7}$$

$$c(t) = f(t) \odot c(t-1) + i(t) \odot g(W_{cx}x(t) + W_{cm}m(t-1) + b_c) \tag{8}$$

$$o(t) = \sigma(W_{ox}x(t) + W_{om}m(t-1) + W_{oc}c(t) + b_o) \tag{9}$$

$$m(t) = o(t) \odot h(c(t)) \tag{10}$$

$$VP(t) = (W_{ym}m(t) + b_y) \tag{11}$$

Table 2 Parameters of the hybrid CNN-LSTM model

Parameter	Value
Number of convolutional filters	132
Kernel size of convolutional layer	1
Kernel size of pooling layer	1
Number of pooling layers	1
Number of LSTM neurons	50
Number of epochs	100
Batch size	10
Training approach for parameters	Adam
Loss function	MAE

3.3 Performance evaluation metrics

The results of the CNN-LSTM prediction model is evaluated based on MAPE, MAE, MSE, and R^2 metrics which are widely used in assessing the performance of time series prediction problem. The metrics are expressed in Eqs. 12-15.

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{12}$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}|y_t - \hat{y}_t| \tag{13}$$

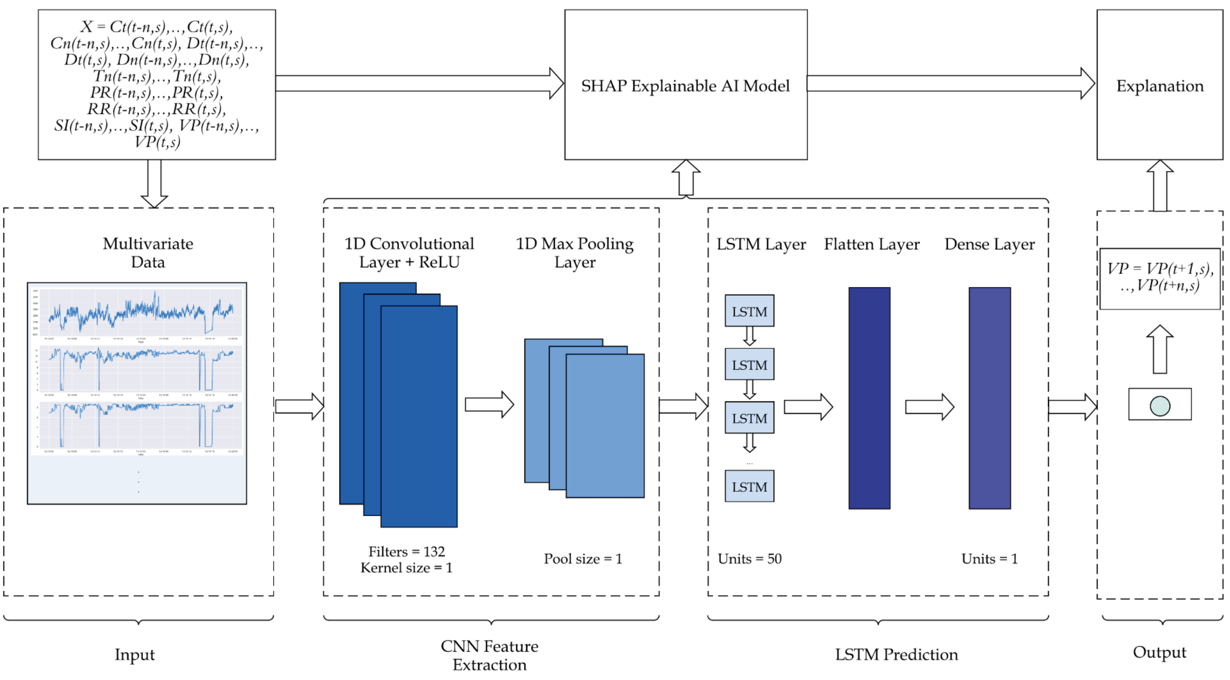


Fig. 4 CNN-LSTM model architecture

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(\hat{y}_t - y_t)^2 \tag{14}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{N}(y_t - \bar{y}_t)^2} \tag{15}$$

where N is the sample size of the test set, $\hat{y}_t$ represents the predicted value of VP at time t, y is the mean value of VP, and y_t represents the observed value of VP at time t.

3.4 SHAP XAI model

Deep neural networks' interpretability has always been a barrier in explaining characteristics used in modeling, especially in many healthcare applications. However, the model interpretation is critical in healthcare applications due to the desire to provide urgent and vital information directly to doctors and patients. Therefore, people have been reluctant to use deep learning-based methods and prefer structural models that produce more straightforward results to be understood. Among the machine learning models for healthcare applications, tree-based methods are preferred for their capability of providing explainable structures and results. In order to obtain the high accuracy provided by the deep learning while inheriting the model interpretability, an XAI mechanism is integrated in the proposed CNN-LSTM-based prediction framework.

A common approach to achieve interpretability is by applying interpretable surrogate models between the inputs and outputs of the neural network to approximate the predictions of the underlying model as accurately as possible and to be interpretable at the same time. Examples of such models can be found as Local Interpretable Model-Agnostic Explanations (LIME) [64] and Shapley Additive Explanations (SHAP) [65]. LIME implements an interpretation of an individual instance's prediction based on a local surrogate model while SHAP provides both local and global explanations for its underlying model based on Shapley values that describe a way to distribute the total gains to all features by assuming they all collaborate. For vaccine prediction, it is important to derive both global explanations and local explanations. Global explanations are desirable for vaccine manufacturers to identify explanations of vaccine demand needs over long periods of time instead of daily need to better plan vaccine manufacturing and supply. Therefore, SHAP is used in this work to explain the prediction results of the CNN-LSTM model. Shapley value was initially developed for game theory in order to determine the importance of each player in a multiplayer game. The Shapley value indicates the value of a player across all combinations of players present in the game. In recent years, the SHAP method has been used to explain machine learning and deep learning models [66]. For vaccine rate prediction, a Shapley value is assigned to each feature showing its importance for prediction and explaining how much each feature contributes to the prediction results.

The SHAP method is used to explain the results of the CNN-LSTM model and identify the features contributing to the prediction results for a single instance and across all instances. This involves training the CNN-LSTM model on all the features. Each feature's contribution to the model is underlined by the Shapley value ϕ_i as calculated with Eq. 16, where S_x represents a set of all features, and f represents the characteristic function used to estimate the utility of a coalition. $f_S(x_S)$ is the coefficient of determination obtained after training the model on the features x_S in subset S [65]. The Shapley values help explain the prediction results for vaccination rates and the features affecting the prediction results, removing the black-box nature of the deep learning model. For instance, the vaccination rate increase for a given day can be explained by the increase in features such as the number of confirmed cases or the reproduction rate of the disease. The overall features' effect on the vaccination prediction can also be described as part of global interpretability to specify which features had the highest impact in driving the vaccination rate either upward or downward. The knowledge obtained by identifying features contributing to vaccination rates can be critical to decision-makers for vaccine manufacturing and supply planning. However, it is important to note that explanations provided by SHAP cannot provide causal relationships and thus need to be interpreted cautiously.

$$\phi_i = \sum_{S\subseteq S_x\setminus\{i\}} \frac{|S|!(|S_x| - |S| - 1)!}{|S_x|!}\left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right] \tag{16}$$

4 Experimental results

4.1 Case studies

Three case studies were conducted involving low-income developing countries from different continents to ensure the generalizability of the proposed model across countries with diverse geographical environments. Nigeria, Nepal, and Peru were selected for vaccination rate prediction from Africa, Asia, and South America continents representing diverse locations. In addition, these countries from different continents started the vaccination comparatively later than other countries due to issues with vaccine supply. They represent countries with low income and limited resources where vaccine shortage or oversupply can significantly impact the health of the population and are therefore of greater importance to accurately predict their vaccination rate and anticipate their vaccine demand in advance. At the beginning of vaccine distribution worldwide, such countries suffered from a shortage of vaccines due to their high population and insufficient vaccine availability. After a vaccination plan was set and vaccine distribution started, they suffered from oversupply, where millions of vaccines were disposed of due to the hesitancy of the population to take vaccines fueled by disbelief from social media. Therefore, for countries with low income, it is essential to accurately predict vaccination rates to plan and prepare vaccine supply in advance and reduce the potential shortage or waste. Shortage of supply may result in an increased number of cases and consequently, for a country with limited resources, an increase in the number of deaths. On the other hand, oversupply may cause vaccine expiration, especially with vaccines being considered perishable items and not having enough available storage with special treatment to store them. In these case studies, COVID-19 vaccination records, confirmed cases, and testing datasets of Nigeria, Nepal, and Peru from the start of vaccination are used [67]. The vaccination rate is set as the target feature and the historical vaccination rates along with the selected features are used to predict the vaccination rates on a weekly, bi-weekly, and monthly basis. The dataset is represented as a time series data where for each day the vaccination and potential features data are identified. A total of 359 days of vaccination data were used in the case studies for training and testing the model depending on when each country initiated the vaccination. In addition, a total of 359 days for 34 features including vaccination, testing, and confirmed cases and deaths data were taken from the start of vaccination and considered as candidate features as shown in Appendix A.

4.2 Extracted and constructed features

All features extracted from Nigeria, Nepal, and Peru's datasets are first cleaned and normalized. Then, two additional features, $RR(t, s)$ and $SI(t, s)$, are constructed based on domain knowledge. Figure 5 shows the $RR(t, s)$ calculated for Nigeria for the given time period. As $RR(t, s)$ indicates the potential spread of the disease, a value > 1 indicates that the infection may spread in the population while a value < 1 indicates that the disease will not spread in a population.

The new feature calculated by the government response tracker, $SI(t, s)$, combines information from nine metrics regarding the government response to COVID-19 in terms of restrictions. It can potentially avoid the overfitting problem with all nine metrics included while maintaining sufficient information in one feature for prediction. Figure 6 shows the $SI(t, s)$ for Nigeria for the time period taken. Generally, a lower $SI(t, s)$ value indicates less strict government policies, making the disease spread more easily in that period of time. This gives new knowledge about relation between the disease and policies. The extracted and constructed features are preprocessed to ensure all features are cleaned and normalized for the next phase.

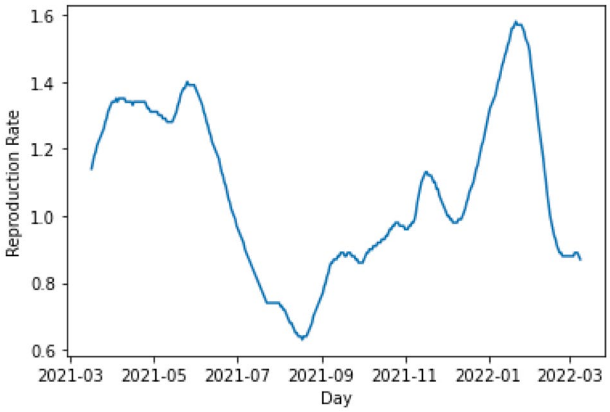


Fig. 5 Nigeria's Reproduction Rate from March 16th, 2021 to March 9th, 2022

4.3 Selected features

The objective of feature selection based on filtering and wrapping is to identify features that have the highest impact on vaccination rates and improve its prediction accuracy. All the extracted features from the original datasets provided in Appendix A and the two constructed features are first fed into the filter method of Pearson correlation and mutual information. Then, a wrapper method of RFE is performed on the candidate features with a threshold set to 8 considering the computational efficiency. The RFE is performed

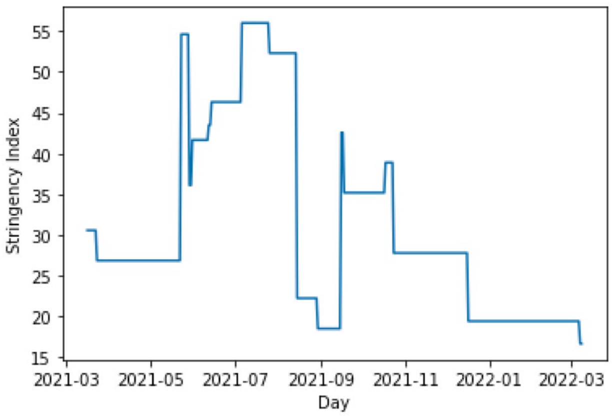


Fig. 6 Nigeria's Stringency Index from March 16th, 2021 to March 9th, 2022

Table 3 Selected features for Nigeria

Annotation	Feature	Description
Ct	Total cases	Total confirmed cases of COVID-19
Cn	New cases	New confirmed cases of COVID-19
Dt	Total deaths	Total deaths attributed to COVID-19
Dn	New deaths	New deaths attributed to COVID-19
Tn	New tests	New tests for COVID-19
PR	Positive rate	The share of COVID-19 tests that are positive
RR	Reproduction rate	Estimate of the effective RR of COVID-19
SI	Stringency index	Government response SI
VP	Vaccinated people per hundred in the population	Total number of people who received at least one vaccine dose per 100 people in the total population

using three models, including Random Forest, Decision Tree, and XGBoost, which are widely used to verify the effect of the selected features on the target feature. The selected features identified when applying the three models are the same. Tables 3, 4, and 5 describe the selected features for Nigeria, Nepal, and Peru, respectively, after filter and wrapper methods are applied, along with the historical vaccination rates. The Pearson correlation and feature importance using Random Forest of selected features for Nigeria are provided in Figs. 7 and 8 as references. It is observed that the total cases (Ct) and the total deaths (Dt) are the highest contributing features that affect the vaccination rates for Nigeria.

4.4 Prediction and performance comparison

To verify the impact of the enhanced feature engineering process and the effectiveness of the proposed CNN-LSTM prediction model, its performance is compared to the traditional statistical forecasting model ARIMA and the standalone univariate and multivariate LSTM models. The CNN-LSTM prediction model was validated using cross-validation on a rolling basis with three folds, where 179 days are used for training and 60 days for testing in the first fold, the 239 days of data are used for training and testing in the first fold are then used for training in the second fold and another 60 days are used for testing. For the third fold, the training and testing data used in the second fold of 299 days is used for training and the remaining data of 60 days is used for testing. The proposed model achieved the lowest MAPE compared to other models for weekly, bi-weekly, and monthly prediction. MAPE is used for comparison since it is scale-independent and can be easily interpreted as the average difference between the predicted value and the actual value and is expressed with a percentage. The comparison between univariate and multivariate LSTM models indicates that adding the selected features by utilizing domain knowledge with the enhanced feature engineering process reduced the MAPE by an average of 11%, 32%, and 47% for Nigeria, Nepal, and Peru, respectively, for weekly, bi-weekly, and monthly prediction, indicating the effectiveness of the feature engineering process. In addition, integrating the CNN to the LSTM model for multivariate prediction reduced the MAPE further by an average of 13%, 11%, and 23% for Nigeria, Nepal, and Peru, respectively, for weekly, bi-weekly, and monthly prediction. The results of the comparison between the models are summarized in Tables 6, 7, and 8. Besides the general

Table 4 Selected features for Nepal

Annotation	Feature	Description
Ct	Total cases	Total confirmed cases of COVID-19
Dtm	Total deaths per million	Total deaths attributed to COVID-19 per 1,000,000 people
Ttt	Total tests per thousand	Total tests for COVID-19 per 1,000 people
Ctm	Total cases per million	Total confirmed cases of COVID-19 per 1,000,000 million
Dt	Total deaths	Total deaths attributed to COVID-19
Dns	New deaths smoothed	New deaths attributed to COVID-19 (7-day smoothed)
PR	Positive rate	The share of COVID-19 tests that are positive, given as a rolling 7-day average
Cnsm	New cases smoothed per million	New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people
VP	Vaccinated people per hundred in the population	Total number of people who received at least one vaccine dose per 100 people in the total population

Table 5 Selected features for Peru

Annotation	Feature	Description
Dtm	Total deaths per million	Total deaths attributed to COVID-19 per 1,000,000 people
Cns	New cases smoothed	New confirmed cases of COVID-19 (7-day smoothed)
Ct	Total cases	Total confirmed cases of COVID-19
Dt	Total deaths	Total deaths attributed to COVID-19
Vt	Total vaccinations	Total number of COVID-19 vaccination doses administered
PR	Positive rate	The share of COVID-19 tests that are positive
Cnsm	New cases smoothed per million	New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people
Ctm	Total cases per million	Total confirmed cases of COVID-19 per 1,000,000 million
VP	Vaccinated people per hundred in the population	Total number of people who received at least one vaccine dose per 100 people in the total population

prediction accuracy comparison, the performance of different models is also evaluated by comparing them to the actual vaccination rates time series for the 60 days of testing period when using 299 days for training. The weekly prediction results for Nigeria's models are shown in Fig. 9. The prediction results can help the Center of Disease and Control (CDC) identify when it will achieve its target of vaccinating 70% of the population to achieve herd immunity. Moreover, the vaccination rates identified by the prediction models are translated to the number of vaccines needed considering the country's population by assuming that each individual will receive one vaccine. For developing countries, the priority is to ensure a high percentage of the population is vaccinated. Therefore, the anticipated demand to ensure a high percentage of the population is vaccinated assumes that each individual will receive one vaccine dose and does not consider second doses and boosters. The results for Nigeria's vaccine demand are shown in Fig. 10. It is observed that the gap of vaccine demand using different prediction models ranges from around one hundred to two hundred thousand, which signifies the importance of using the CNN-LSTM model to identify vaccine demand. Based on the vaccine demand results, the government can plan the vaccine supply accordingly in advance by considering a prediction interval based on the prediction results.

4.5 Interpretable CNN-LSTM model

The results of the CNN-LSTM model in predicting vaccination rates are interpreted through the SHAP method to identify each factor's effect on the vaccination rates for each country. The contribution of each factor to the vaccination rates is analyzed to determine if it has a positive or negative effect on the predicted vaccination rates. It derives an understanding of future vaccination rates and vaccine demand for assisting government and vaccine

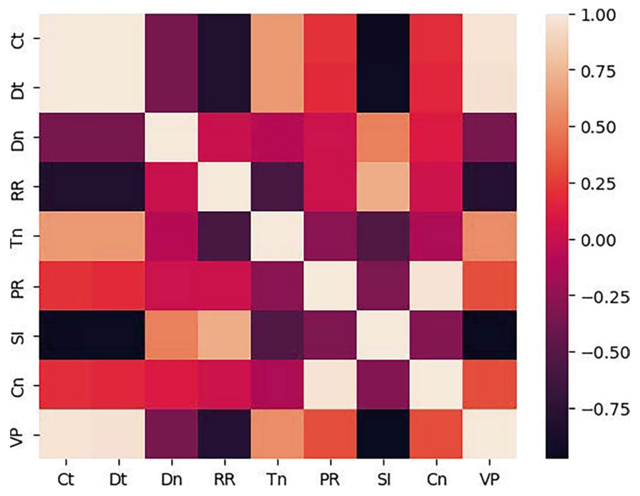


Fig. 7 Pearson Correlation for selected features for Nigeria

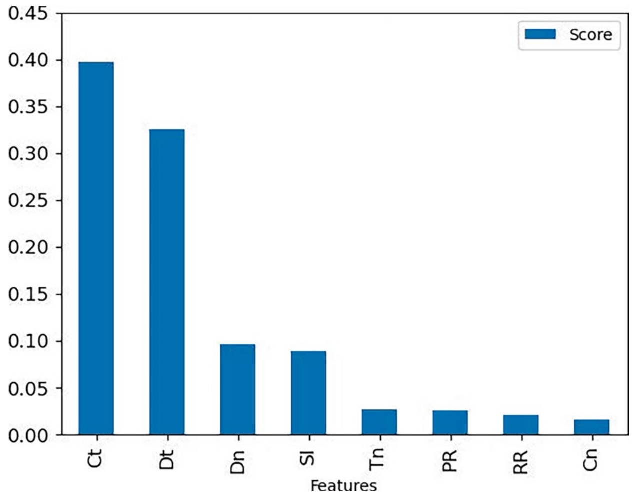


Fig. 8 Random Forest feature importance for Nigeria

Table 6 Model comparison for Nigeria

	Metric	ARIMA	Univariate LSTM	Multivariate LSTM	**Multivariate CNN-LSTM**
Weekly Prediction	MAPE	24.9221	13.1660	11.9818	**09.9920**
	MAE	00.1450	00.0404	00.0347	00.0299
	MSE	00.0200	00.0047	00.0036	00.0035
	R^2	00.0522	00.8321	00.8686	00.8818
Bi-weekly Prediction	MAPE	24.9042	12.5979	11.6276	**09.8306**
	MAE	00.1460	00.0381	00.0339	00.0303
	MSE	00.0200	00.0042	00.0035	00.0036
	R^2	00.1402	00.8448	00.8789	00.8774
Monthly Prediction	MAPE	24.8369	13.3379	11.1675	**10.2827**
	MAE	00.1460	00.0389	00.0345	00.0299
	MSE	00.0200	00.0042	00.0040	00.0034
	R^2	00.2982	00.8376	00.8684	00.8827

Table 7 Model comparison for Nepal

	Metric	ARIMA	Univariate LSTM	Multivariate LSTM	**Multivariate CNN-LSTM**
Weekly Prediction	MAPE	29.9910	18.9221	11.9002	**10.7501**
	MAE	00.1003	00.0547	00.0263	00.0252
	MSE	00.0134	00.0063	00.0012	00.0012
	R^2	00.4794	00.7728	00.9608	00.9535
Bi-weekly Prediction	MAPE	28.0328	18.8527	12.9485	**11.9420**
	MAE	00.0927	00.0530	00.0282	00.0277
	MSE	00.0122	00.0059	00.0014	00.0014
	R^2	00.3478	00.8033	00.9539	00.9500
Monthly Prediction	MAPE	27.2672	19.7232	14.1931	**12.1100**
	MAE	00.0889	00.0598	00.0293	00.0261
	MSE	00.0115	00.0090	00.0015	00.0014
	R^2	00.2700	00.7412	00.9510	00.9537

Table 8 Model comparison for Peru

	Metric	ARIMA	Univariate LSTM	Multivariate LSTM	**Multivariate CNN-LSTM**
Weekly Prediction	MAPE	80.4632	19.0320	11.1930	**08.7241**
	MAE	00.0846	00.0486	00.0229	00.0187
	MSE	00.0092	00.0050	00.0008	00.0006
	R^2	00.0488	00.8300	00.9565	00.9708
Bi-weekly Prediction	MAPE	79.7342	20.0328	10.3487	**08.6407**
	MAE	00.0838	00.0521	00.0210	00.0180
	MSE	00.0088	00.0058	00.0007	00.0005
	R^2	00.0049	00.7797	00.9637	00.9708
Monthly Prediction	MAPE	80.5447	23.3810	11.3630	**07.8095**
	MAE	00.0851	00.0598	00.0280	00.0180
	MSE	00.0089	00.0080	00.0008	00.0005
	R^2	00.0144	00.7202	00.9543	00.9718

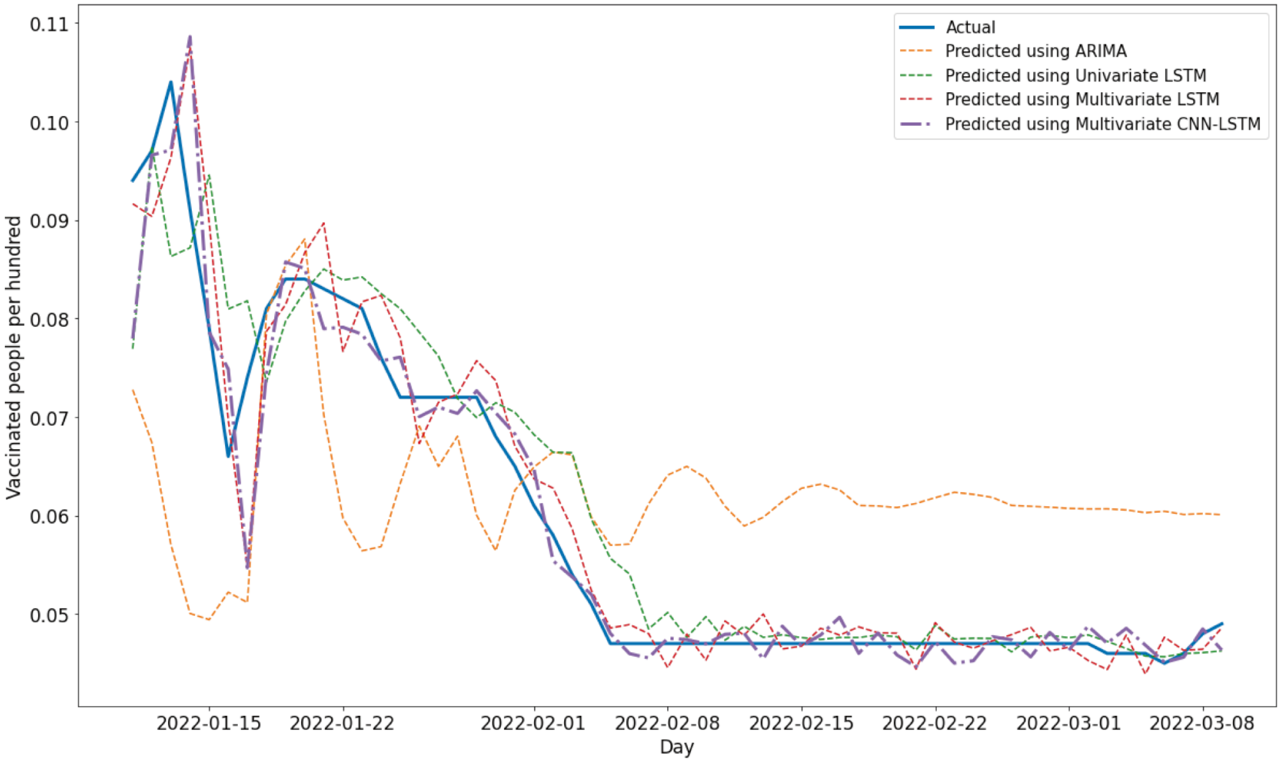


Fig. 9 Weekly prediction of vaccinated people in Nigeria

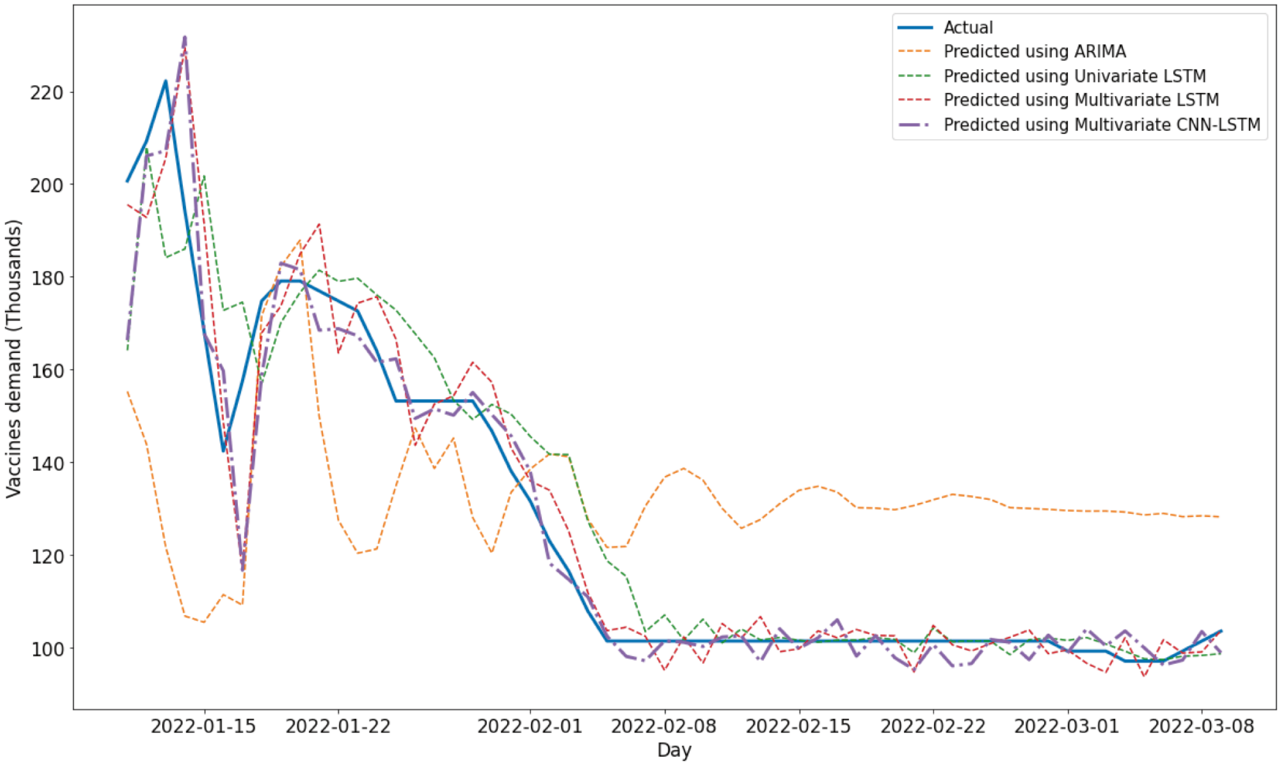


Fig. 10 Weekly vaccine demand in Nigeria

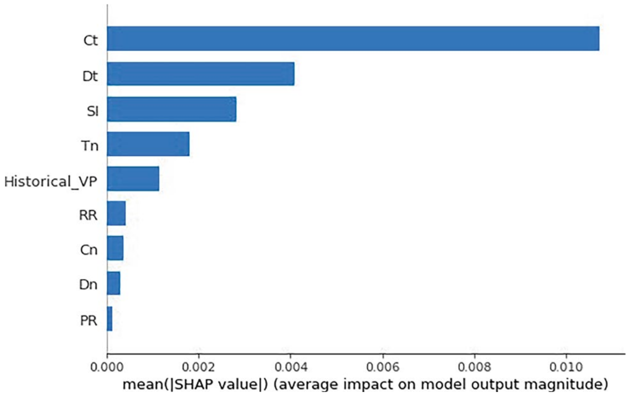
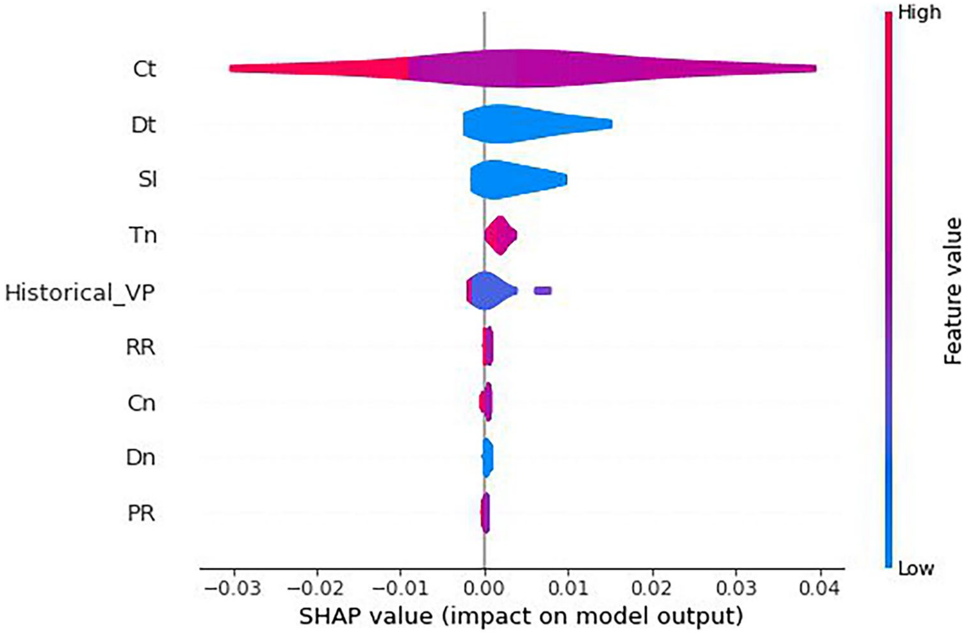


Fig. 11 SHAP feature importance for Nigeria vaccination rate prediction

manufacturers with decision-making. Take Nigeria as an example. Figure 11 shows the SHAP feature importance for Nigeria, where the feature with the highest impact is the total confirmed cases followed by total deaths. It is aligned with the feature selection results provided by the Random Forest, further emphasizing the effectiveness of the feature engineering process. Figure 12 shows the average contribution of the factors considered on the weekly predicted vaccination rates for Nigeria. One noticeable point is that a larger number of confirmed cases results in lower future vaccination rates. It can be explained by people's hesitancy due to their observation of a large number of positive cases and their belief that vaccination is not effective. On the contrary, a lower number of total deaths due to COVID-19 can encourage more people to take vaccination as indicated by the higher future vaccination rates. In addition, lower SI for showing the less strict government policies also positively affects the future vaccination rates. If the government lifts the restrictions on travel and gatherings, and reduces the number of campaigns, it may convey the effectiveness of the vaccines in controlling the pandemic to the public. Similar interpretation can be done for the other two countries based on the SHAP feature analysis. Here we only illustrate the conclusion. For Nepal, the total cases and total deaths per million are identified as the features with the highest impact on the vaccination rates. Similar to the situation in Nigeria, high death rate rather than the large number of positive cases is the main motivation for people to take vaccine. For Peru, the historical vaccination rates is the top feature affecting the vaccination prediction, indicating the willingness of taking vaccination is highly impacted by the historical and current vaccination status. Different interpretations of features' impact on the future vaccination rates can be derived for different countries and regions to help the government and vaccine manufacturers anticipate the future increase or decrease of vaccinated people in each country or region. The insights provided by the SHAP method extend on the typical deep learning method and make it more visible and clear to interpret the effect of features on the prediction results.

Moreover, besides the global interpretability the SHAP method provides, it also considers local interpretability. The SHAP method shows the important features affecting the prediction results for a certain day. For instance, Fig. 13 shows the factors affecting the prediction results for a random testing day for Nigeria. The base value for the prediction is around 0.395, which indicates that the

Fig. 12 SHAP value impact for Nigeria



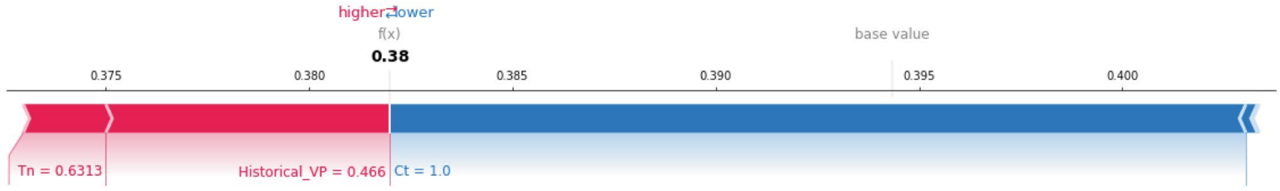


Fig. 13 SHAP instance interpretation plot of the CNN-LSTM prediction result for Nigeria

expected predicted vaccination rate without considering the features is 0.395 for this specific day. The actual predicted vaccination rate 0.380 is obtained by considering the features' contribution. The difference between the base value and the actual predicted value is around 0.015, indicating the features' effect. Specifically, the number of total cases (Ct) leads the prediction downward to a lower vaccination rate compared to the base value. It also has the highest impact on the vaccination rate for this day compared to other features. The new tests (Tn) and historical vaccination rates (Historical VP) drive the predicted vaccination rate to a higher level closer to the base value but it fails to reach the base value because of the magnitude of Ct in driving the prediction downward. The values for each of these features, such as Tn=0.6313, represent the normalized feature value contributing to the prediction. Since the features are all normalized, the impact indicated in Fig. 13 is independent of the scale as well.

5 Conclusion

Explainable prediction of a country's COVID-19 vaccination rates is essential for government and vaccine manufacturers to identify vaccine demand and better plan its manufacturing, supply, and distribution and provide herd immunity. Previous studies have proposed different methods to predict future vaccination rates according to historical vaccination records. However, the impact of positive cases, deaths, testing, and domain knowledge was not studied and incorporated into the prediction. In addition, there is no sufficient discussion on applying explainable deep-learning-based methods in predicting the COVID-19 vaccination rates based on multiple features and interpreting the impact of each feature accordingly. The proposed hybrid CNN-LSTM model, to the best of our knowledge, is the first attempt that integrates the critical features produced by a three-stage feature engineering process in predicting the COVID-19 vaccination rate. To achieve the interpretability of the proposed hybrid CNN-LSTM model, an XAI method based on the SHAP is developed to further explain and understand the effect of each feature on the vaccination rate. In the case of vaccination rates, the model interpretability is an important aspect that should be added to improve model transparency for extracting key information for decision-making.

Based on case studies on Nigeria, Nepal, and Peru, a hybrid CNN-LSTM model is developed and applied to predict the vaccination rate on a weekly, bi-weekly, and monthly basis. To further evaluate its prediction performance, it is compared to three benchmark models including statistical model ARIMA, univariate LSTM, and multivariate LSTM, and validated using cross-validation on a rolling basis. The results show that a significant improvement with an average 30% decrease in MAPE is achieved by introducing the selected features. An additional 16% improvement is provided by CNN, where a new feature space is constructed to digest the important information nested in the original features before entering the LSTM to learn the temporal dependence among or within features.

In the future, the proposed framework will be further extended to improve the prediction performance by optimizing the hyperparameters in the feature selection and neural network characteristics. In addition, the response indicators used in calculating the stringency index will be further explored to identify how individual policies such as public information campaigns will affect future vaccine prediction. Different spatial factors will be considered besides the government policies of the countries such as the geographical locations of the countries. Moreover, an ongoing study based on this work will discuss a deep-learning-based model to transfer knowledge from one domain to another to overcome insufficient training due to the limited data. In the prediction of vaccination rates or other medical product demand, it is common in practice that only limited data is available because of the sudden outbreaks of a pandemic in underdeveloped regions. Therefore, the model can be further improved to learn from the knowledge of one disease or a region to predict the demand for the target region or a new respiratory infectious disease with limited available data.

Appendix A: list of features

Table 9 COVID-19 vaccination data

Annotation	Feature	Description
VP	New people vaccinated smoothed per hundred	Daily number of people receiving their first vaccine dose (7-day smoothed) per 100 people in the total population
Vt	Total vaccinations	Total number of COVID-19 vaccination doses administered
PV	People vaccinated	Total number of people who received at least one vaccine dose
PFV	People fully vaccinated	Total number of people who received all doses prescribed by the initial vaccination protocol
Vn	New vaccinations	New COVID-19 vaccination doses administered
Vns	New vaccinations smoothed	New COVID-19 vaccination doses administered (7-day smoothed)
Vth	Total vaccinations per hundred	Total number of COVID-19 vaccination doses administered per 100 people in the total population
PFVh	People fully vaccinated per hundred	Total number of people who received at least one vaccine dose per 100 people in the total population
Vnsm	New vaccinations smoothed per million	New COVID-19 vaccination doses administered (7-day smoothed) per 1,000,000 people in the total population
Bt	Total boosters	Total number of COVID-19 vaccination booster doses administered
PVh	People vaccinated per hundred	Total number of people who received at least one vaccine dose per 100 people in the total population
Bth	Total boosters per hundred	Total number of COVID-19 vaccination booster doses administered per 100 people in the total population
PVn	New people vaccinated smoothed	Daily number of people receiving their first vaccine dose (7-day smoothed)

Table 10 COVID-19 testing data

Annotation	Feature	Description
Tt	Total tests	Total tests for COVID-19
Tn	New tests	New tests for COVID-19
Ttt	Total tests per thousand	Total tests for COVID-19 per 1,000 people
Tns	New tests smoothed	New tests for COVID-19 (7-day smoothed)
Tnst	New tests smoothed per thousand	New tests for COVID-19 (7-day smoothed) per 1,000 people
PR	Positive rate	The share of COVID-19 tests that are positive, given as a rolling 7-day average
Tc	Tests per case	Tests conducted per new confirmed case of COVID-19, given as a rolling 7-day average

Table 11 COVID-19 confirmed cases and deaths data

Annotation	Feature	Description
Ct	Total cases	Total confirmed cases of COVID-19
Cn	New cases	New confirmed cases of COVID-19
Cns	New cases smoothed	New confirmed cases of COVID-19 (7-day smoothed)
Dt	Total deaths	Total deaths attributed to COVID-19
Dn	New deaths	New deaths attributed to COVID-19
Dns	New deaths smoothed	New deaths attributed to COVID-19 (7-day smoothed)
Ctm	Total cases per million	Total confirmed cases of COVID-19 per 1,000,000 people
Cnm	New cases per million	New confirmed cases of COVID-19 per 1,000,000 people
Cnsm	New cases smoothed per million	New confirmed cases of COVID-19 (7-day smoothed) per 1,000,000 people
Dtm	Total deaths per million	Total deaths attributed to COVID-19 per 1,000,000 people
Dnm	New deaths per million	New deaths attributed to COVID-19 per 1,000,000 people
Dnsm	New deaths smoothed per million	New deaths attributed to COVID-19 (7-day smoothed) per 1,000,000 people

Author contributions Lubna Altarawneh: Conceptualization, Investigation, Programming, Writing - Original Draft, Writing - Review & Editing, Hao Wang: Programming Yu Jin: Conceptualization, Supervision, Writing - Review & Editing.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Availability of data Data is available at https://github.com/owid/covid-19-data/tree/master/public/data.

Code availability Not Applicable.

Declarations

Ethics approval Not Applicable.

Consent to participate Not Applicable.

Consent for publication Not Applicable.

Conflict of interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

References

1. Sim SW, Moey KSP, Tan NC. The use of facemasks to prevent respiratory infection: a literature review in the context of the health belief model. Singapore Med J. 2014;55(3):160.
2. Choi Y, Kim JS, Kim JE, Choi H, Lee CH. Vaccination prioritization strategies for covid-19 in korea: a mathematical modeling approach. Int J Environ Res Public Health. 2021;18(8):4240.
3. Zhou X, Li Y. Forecasting the covid-19 vaccine uptake rate: an infodemiological study in the us. Hum Vacc Immunother. 2022;18(1):2017216.
4. Tamilselvan S, Kumar S, Priya J, Ramesh R. Covid vaccination analysis and prediction using machine learning. In Smart Technologies, Communication and Robotics (STCR), 2021; pp. 1–5. IEEE.
5. Kane MJ, Price N, Scotch M, Rabinowitz P. Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. BMC Bioinform. 2014;15(1):1–9.
6. 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), 2018; pp. 1394–1401. IEEE.
7. Wang W, Mao W, Tong X, Gang X. A novel recursive model based on a convolutional long short-term memory neural network for air pollution prediction. Remote Sens. 2021;13(7):1284.
8. Kühnert C, Gonuguntla NM, Krieg H, Nowak D, Thomas JA. Application of lstm networks for water demand prediction in optimal pump control. Water. 2021;13(5):644.
9. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
10. Zhang X, Mahadevan S. Bayesian neural networks for flight trajectory prediction and safety assessment. Decis Support Syst. 2020;131:113246.
11. Xu J, Rahmatizadeh R, Bölöni L, Turgut D. A sequence learning model with recurrent neural networks for taxi demand prediction. In 2017 IEEE 42nd Conference on Local Computer Networks (LCN), 2017; pp. 261–268. IEEE.
12. Patol D, Mehta R, Mehta C, Naik G, Agrawal H. Taxi demand prediction using lstm and optimized taxi geo-distribution. In Soft Computing and Signal Processing, 2022; pp 163–172. Springer.
13. Niu K, Wang C, Zhou X, Zhou T. Predicting ride-hailing service demand via rpa-lstm. IEEE Trans Veh Technol. 2019;68(5):4213–22.
14. Afrin T, Yodo N. A lstm-based correlated traffic data prediction framework. Knowl Based Syst. 2022;237:107755.
15. Yun KK, Yoon SW, Won D. Prediction of stock price direction using a hybrid ga-xgboost algorithm with a three-stage feature engineering process. Expert Syst Appl. 2021;186:115716.
16. Sahisnu JS, Natalia F, Ferdinand FV, Sudirman S, Ko CS. Vaccine prediction system using arima method. ICIC Express Lett B Appl. 2020;11(6):567–575.
17. Kim K-W, Li G, Park S-T, Ko M-H. A study on birth prediction and bcg vaccine demand prediction using arima analysis. Indian J Sci Technol. 2016;9(24):1–7.
18. Arifin B, Anas T. Lessons learned from covid-19 vaccination in indonesia: experiences, challenges, and opportunities. Hum Vacc Immunother. 2021;17(11):3898–906.
19. Zheng Y-J, Chen X, Song Q, Yang J, Wang L. Evolutionary optimization of covid-19 vaccine distribution with evolutionary demands. IEEE Trans Evol Comput. 2022.
20. Rathinasamy M, Kandhasamy S. An exploratory analysis of covid-19 vaccination administration. Proc Indian National Sci Acad. 2022;88(1):43–52.
21. Yamak PT, Yujian L, Gadosey PK. A comparison between arima, lstm, and gru for time series forecasting. In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence. 2019; pp. 49–55.
22. Rathnayaka RK, Seneviratna DM, Jianguo W, Arumawadu HI. A hybrid statistical approach for stock market forecasting based on artificial neural network and arima time series models. In 2015 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC), 2015; pp. 54–60. IEEE.
23. Ribeiro MHDM, da Silva RG, Mariani VC, dos Santos Coelho L. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos Solit Fractals. 2020;135:109853. Elsevier.
24. Tamang SK, Singh PD, Datta B. Forecasting of Covid-19 cases based on prediction using artificial neural network curve fitting technique. Global J Environ Sci Manag. 2020;6(Covid-19):53–64. GJESM Publisher.
25. Somyanonthanakul R, Warin K, Amasiri W, Mairiang K, Mingmalairak C, Panichkitkosolkul W, Silanun K, Theeramunkong T, Nitikraipot S, Suebnukarn S. Forecasting COVID-19 cases using time series modeling and association rule mining. BMC Med Res Methodol. 2022;22(1):281. Springer.
26. Chen X, Huang H, Jiandong J, Sun R, Zhang J. Impact of vaccination on the covid-19 pandemic in us states. Sci Rep. 2022;12(1):1–10.
27. Coccia M. Optimal levels of vaccination to reduce covid-19 infected individuals and deaths: a global analysis. Environ Res. 2022;204:112314.
28. Dong E, Ratcliff J, Goyea TD, Katz A, Lau R, Ng TK, Garcia B, Bolt E, Prata S, Zhang D, et al. The johns hopkins university center for systems science and engineering covid-19 dashboard: data collection process, challenges faced, and lessons learned. Lancet Infect Dis. 2022.
29. Hasell J, et al. A cross-country database of covid-19 testing. Sci Data. 2020;7(1):345.
30. Gabler J, Raabe T, Röhrl K, von Gaudecker H-M. The effectiveness of testing, vaccinations and contact restrictions for containing the covid-19 pandemic. Sci Rep. 2022;12(1):1–8.
31. Said AB, Erradi A, Aly HA, Mohamed A. Predicting covid-19 cases using bidirectional lstm on multivariate time series. Environ Sci Pollut Res. 2021;28(40):56043–56052.

32. Borges D, Nascimento MCV. Covid-19 icu demand forecasting: a two-stage prophet-lstm approach. Appl Soft Comput. 2022;125:109181.
33. Das S. Filters, wrappers and a boosting-based hybrid for feature selection. In Icml. 2001;1:74–81. Citeseer.
34. Sen S, Saha S, Chatterjee S, Mirjalili S, Sarkar R. A bi-stage feature selection approach for covid-19 prediction using chest ct images. Appl Intell. 2021;51(12):8985–9000.
35. Ali RH, Abdulsalam WH. The prediction of covid-19 disease using feature selection techniques. J Phys Conf Series. 2021;1879:022083. IOP Publishing.
36. Saha A, Mishra V, Rath SK. Prediction of growth in covid-19 cases in india based on machine learning techniques. In 2022 International Conference on Innovative Trends in Information Technology (ICITIIT), 2022; pp. 1–6. IEEE.
37. Pourhomayoun M, Shakibi M. Predicting mortality risk in patients with covid-19 using machine learning to help medical decision-making. Smart Health. 2021;20:100178.
38. Silva L. A feature engineering approach to wind power forecasting: Gefcom 2012. Int J Forecast. 2014;30(2):395–401.
39. Zheng H, Yanghui W. A xgboost model with weather similarity analysis and feature engineering for short-term wind power forecasting. Appl Sci. 2019;9(15):3019.
40. Pham T, Tao X, Zhang J, Yong J, Li Y, Xie H. Graph-based multi-label disease prediction model learning from medical data and domain knowledge. Knowl Based Syst. 2022;235:107662.
41. Yue W, Yunjie X, Li J. Feature construction for fraudulent credit card cash-out detection. Decis Support Syst. 2019;127:113155.
42. Chandra R, Jain A, Chauhan DS. Deep learning via lstm models for covid-19 infection forecasting in India. PloS One. 2022;17(1):e0262708.
43. Chen S, Paul R, Janies D, Murphy K, Feng T, Thill J-C. Exploring feasibility of multivariate deep learning models in predicting covid-19 epidemic. Front Public Health. 2021;9:765.
44. Wathore R, Rawlekar S, Anjum S, Gupta A, Bherwani H, Labhasetwar N, Kumar R. Improving performance of deep learning predictive models for covid-19 by incorporating environmental parameters. Gondwana Res. 2022.
45. Widiputra H. Ga-optimized multivariate cnn-lstm model for predicting multi-channel mobility in the covid-19 pandemic. Emerg Sci J. 2021;5(5):619–35.
46. Zain ZM, Alturki NM. Covid-19 pandemic forecasting using cnn-lstm: a hybrid approach. J Control Sci Eng. 2021;2021.
47. Park S, Yang J-S. Interpretable deep learning lstm model for intelligent economic decision-making. Knowl Based Syst. 2022;248:108907.
48. Srinivasu PN, Sandhya N, Jhaveri RH, Raut R. From blackbox to explainable ai in healthcare: existing tools and case studies. Mob Inf Syst. 2022;2022.
49. Adadi A, Berrada M. Explainable ai for healthcare: from black box to interpretable models. In Embedded Systems and Artificial Intelligence, 2020; pp. 327–337. Springer.
50. Tascikaraoglu A. Evaluation of spatio-temporal forecasting methods in various smart city applications. Renew Sust Energy Rev. 2018;82:424–35.
51. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature selection: a data perspective. ACM Comput Surv (CSUR). 2017;50(6):1–45.
52. Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Giattino C, Hasell J, Macdonald B, Dattani S, Beltekian D, Ortiz-Ospina E, Roser M. Coronavirus pandemic (covid-19). Our World in Data. 2020. https://ourworldindata.org/coronavirus.
53. Achaiah NC, Subbarajasetty SB, Shetty RM. R0 and re of covid-19: can we predict when the pandemic outbreak will be contained? Indian J Crit Care Med. 2020;24(11):1125.
54. Linka K, Peirlinck M, Kuhl E. The reproduction number of covid-19 and its correlation with public health interventions. Comput Mech. 2020;66(4):1035–50.
55. Arroyo-Marioli F, Bullano F, Kucinskas S, Rondón-Moreno C. Tracking r of covid-19: a new real-time estimation using the kalman filter. PloS One. 2021;16(1):e0244474.
56. Durbin J, Koopman SJ. Time Series Analysis by State Space Methods. 2012;38. OUP Oxford.
57. Zhongyi H, Bao Y, Xiong T, Chiong R. Hybrid filter wrapper feature selection for short-term load forecasting. Eng Appl Artif Intell. 2015;40:17–27.
58. Zhang X, Zhang Q, Chen M, Sun Y, Qin X, Li H. A two-stage feature selection and intelligent fault diagnosis method for rotating machinery using hybrid filter and wrapper method. Neurocomputing. 2018;275:2426–39.
59. Wah YB, Ibrahim N, Hamid HA, Abdul-Rahman S, Fong S. Feature selection methods: case of filter and wrapper approaches for maximising classification accuracy. Pertanika J Sci Technol. 2018;26(1).
60. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90.
61. Xie H, Zhang L, Lim CP. Evolving cnn-lstm models for time series prediction using enhanced grey wolf optimizer. IEEE Access. 2020;8:161519–161541.
62. Livieris IE, Pintelas E, Pintelas P. A cnn-lstm model for gold price time-series forecasting. Neural Comput Appl. 2020;32(23):17351–17360.
63. Kim T-Y, Cho S-B. Predicting residential energy consumption using cnn-lstm neural networks. Energy. 2019;182:72–81.
64. Ribeiro MT, Singh S, Guestrin C. why should i trust you? explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; pp. 1135–1144.
65. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
66. Ancona M, Oztireli C, Gross M. Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In International Conference on Machine Learning, 2019; pp. 272–281. PMLR.
67. Mathieu E, Ritchie H, Ortiz-Ospina E, Roser M, Hasell J, Appel C, Giattino C, Rodés-Guirao L. A global database of covid-19 vaccinations. Nat Hum Behav. 2021;5(7):947–53.