frontiers | Frontiers in Public Health

ORIGINAL RESEARCH
published: 29 July 2022
doi: 10.3389/fpubh.2022.922795

Short-Term Prediction of COVID-19 Using Novel Hybrid Ensemble Empirical Mode Decomposition and Error Trend Seasonal Model

Dost Muhammad Khan[1], Muhammad Ali[1], Nadeem Iqbal[2,3], Umair Khalil[1], Hassan M. Aljohani[4], Amirah Saeed Alharthi[4] and Ahmed Z. Afify[5]**

[1] Department of Statistics, Abdul Wali Khan University Mardan, Mardan, Pakistan, [2] Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, Pakistan, [3] Division of Computer Science, Mathematics and Science, St John's University, New York, NY, United States, [4] Department of Mathematics & Statistics, College of Science, Taif University, Taif, Saudi Arabia, [5] Department of Statistics, Mathematics and Insurance, Benha University, Benha, Egypt

OPEN ACCESS

Edited by:
Muhammad Tahir,
Saudi Electronic University,
Saudi Arabia

Reviewed by:
Sajid Anwar,
Institute of Management Sciences, Pakistan
Abrar Ullah,
Heriot-Watt University,
United Kingdom

***Correspondence:**
Ahmed Z. Afify
ahmed.afify@fcom.bu.edu.eg
Dost Muhammad Khan
dostmuhammad@awkum.edu.pk

Specialty section:
This article was submitted to Public Mental Health, a section of the journal Frontiers in Public Health

Received: 18 April 2022
Accepted: 13 May 2022
Published: 29 July 2022

Citation:
Khan DM, Ali M, Iqbal N, Khalil U, Aljohani HM, Alharthi AS and Afify AZ (2022) Short-Term Prediction of COVID-19 Using Novel Hybrid Ensemble Empirical Mode Decomposition and Error Trend Seasonal Model. Front. Public Health 10:922795. doi: 10.3389/fpubh.2022.922795

In this article, a new hybrid time series model is proposed to predict COVID-19 daily confirmed cases and deaths. Due to the variations and complexity in the data, it is very difficult to predict its future trajectory using linear time series or mathematical models. In this research article, a novel hybrid ensemble empirical mode decomposition and error trend seasonal (EEMD-ETS) model has been developed to forecast the COVID-19 pandemic. The proposed hybrid model decomposes the complex, nonlinear, and nonstationary data into different intrinsic mode functions (IMFs) from low to high frequencies, and a single monotone residue by applying EEMD. The stationarity of each IMF component is checked with the help of the augmented Dicky–Fuller (ADF) test and is then used to build up the EEMD-ETS model, and finally, future predictions have been obtained from the proposed hybrid model. For illustration purposes and to check the performance of the proposed model, four datasets of daily confirmed cases and deaths from COVID-19 in Italy, Germany, the United Kingdom (UK), and France have been used. Similarly, four different statistical metrics, i.e., root mean square error (RMSE), symmetric mean absolute parentage error (sMAPE), mean absolute error (MAE), and mean absolute percentage error (MAPE) have been used for a comparison of different time series models. It is evident from the results that the proposed hybrid EEMD-ETS model outperforms the other time series and machine learning models. Hence, it is worthy to be used as an effective model for the prediction of COVID-19.

Keywords: prediction, COVID-19, ensemble empirical mode decomposition, augmented Dicky-Fuller test, ARIMA, error trend seasonal model

INTRODUCTION

There has been a growing recognition among data analysts and researchers to focus on the prediction of COVID-19 in different parts of the world. The COVID-19 pandemic can be traced back to a group of severe pneumonia cases identified in Wuhan, China, in December 2019 (1). The initial spread of this contagious virus has been linked to a living animal seafood marketplace in Wuhan, pointing to a zoonotic source of the pandemic. However, person-to-person

transmission has driven rapid spread with cumulative numbers reaching 53,164,803 reported cases, and 1,300,576 deaths globally since the start of the pandemic until 14 November 2020 (2). The worst-hit countries are Italy, France, Germany, and United Kingdom (UK) which recorded approximately 5,084,645 reported cases and 151,380 cumulative deaths. Extraordinary measures have been taken by these countries to reduce the viral spread, specifically in the densely populated regions to reduce the chances that sick people might come into contact with healthy ones. In recent times, the prediction of the current pandemic of COVID-19 outbreak is a test for data experts as inadequate information is available on the initial growing curve, and the epidemiological properties of the virus to be fully elucidated. There has been a renewed interest in using time series models to predict the epidemics, namely, SARS, Ebola, influenza, and dengue (3–10). These studies have shown an increasing curiosity in applying time series models as valuable tools in estimating and predicting epidemics. Unlike the regression models that need one response and at least one explanatory variable, univariate time series models are data-driven and can be used for forecasting without any explanatory variables. Predicting the daily confirmed cases and deaths from COVID-19 is hard as compared to the cumulative confirmed cases and deaths. The reason is that the daily data follow a nonlinear and nonstationary pattern, and hence, most of the linear time series models cannot capture its nonstationary characteristics more precisely.

In recent years, the empirical mode decomposition (EMD) and its modified form known as ensemble empirical mode decomposition (EEMD) (11, 12) have emerged as an attractive method for complex signal analysis. Using this method, a complex signal can be partitioned into a limited number of intrinsic mode functions (IMFs), having simpler frequency mechanisms that lead to easy and precise forecasting. The EMD has been extensively applied in numerous areas, such as the investigation of the complex nonlinear sea wave data (13), earthquake data analysis, construction state monitoring (14), diagnosis of faults in the machines (15, 16), prediction of stock markets, exchange rates, and crude oil (17–20).

Commonly, two approaches have been used in the past, i.e., the first one is statistical, and the second one is referred to as a mathematical model for the prediction of different pandemics. Following are a few studies available in the literature that shows the importance of these models to forecast the spread of the pandemics.

The method of serial interval (SI) of the infection was used by Zhao et al. (21) to estimate the value of reproduction rate (R_0). By implementing this method, the estimated value of R_0 for COVID-19 is found to be 2.56 with a 95% confidence band of 249–2.53. Based on the estimated value of R_0, the initial cases of COVID-19 in China followed an exponential growth. The unreported cases from 1 January 2020 to 15 January 2020, are 469 with a 95% prediction interval of 403–540. It is concluded from this study that the unreported cases probably happen during the first 2 weeks of January.

To find out the predicted reproduction number R_0 of COVID-19, the author in Tang et al. (22) used a deterministic compartmental mathematical model with other variables, namely, the progression of the disease, epidemiological status of the individuals, and intervention measures. The method of likelihood has been used for estimation; the estimated control reproduction number was found to be 6.47 with a 95% prediction interval of 5.71 to 7.23. It is also concluded from this study that tracing, isolation, and quarantine can decrease the reproduction and transmission rate of COVID-19.

Three different artificial neural networks (ANN), i.e., multilayer perception (MLP), radial basis function (RBF), and time delay neural network (TDNN) have been compared with the ARIMA model for predicting the hepatitis A virus (HAV) (23). They used 13 years of data on the HAV in Turkey to check the accuracy of ANN and ARIMA models. Based on the smallest values of mean squared error (MSE), normalized mean squared error (NMSE), and mean absolute error (MAE), the method of MLP outperforms other methods to forecast the infections caused by HAV.

A simple mean-field and susceptible-infected-recovered-death (SIRD) model was used by Fanelli et al. (24) to predict the dynamics of the COVID-19 in China, Italy, and France. The simple mean-field model can be used efficiently to find out the time and height of the peak of the cumulative confirmed cases. The peak of the COVID-19 in Italy is around 21 March 2020, with maximum cumulative cases of 2,600. Using the same data for the SIRD model, it is estimated that the recovery rate for the three different countries is the same, whereas rate of death and infection is different.

Different phenomenological models, i.e., the generalized logistic growth model (GLM), Richard growth model, and subepidemic growth models are implemented for the short-term real-time forecast of COVID-19 in the Hubei Province where the virus has been originated for the overall trajectory in China. Among different phenomenological models, the GLM and Richard model yield comparable prediction intervals in Hubei, while the subepidemic model gives a wider interval than the competing models. Furthermore, the prediction intervals obtained by the subepidemic model are much wider than the other two models both in Hubei and other provinces of China (25).

An exponential model has been used to forecast the number of infected people from COVID-19 in Italy (26). Based on the exponent value of $r=0.225$ for the model, the exponential prediction and the actual number of confirmed cases are very much similar. According to this model, the estimated reproduction rate R_0 varies between 2.76 and 3.25, which is very much similar to the one reported initially for the city of Wuhan in China. It is predicted from this model that the cumulative number of confirmed cases in Italy by March 15 will be more than 30,000.

The SIRD model was fitted by Anastassopoulou et al. (27) to estimate the basic reproduction number R_0, daily confirmed cases, and daily deaths along with a 90% confidence interval. Based on the data of the confirmed cases, the average estimated and simulated value of R_0 for the SIRD model is approximately 2.6 and 2. According to this study, the total number of infected people could reach 180,000 with a lower confidence interval of

45,000, and the total number of deceased persons from COVID-19 might be more than 2,700 by February 29. It is also evident from this study that the fatality rates show a declining pattern from January 26.

A simple time series predicting method from the exponential family to forecast the total number of infected people from COVID-19 was used by Petropoulos et al. (28). The forecast accuracy of this method is better than the other time series models and is hence used for short-term forecasting. Models from the exponential family capture both trends and seasonal components based on the nature of the data only trend, and nonseasonal components of the dataset are used in this study. The 10 days ahead forecasted value of cumulative confirmed cases around the globe is 209,000 with a 90% prediction interval from 38,000 to 534,000 in the time window from 01 February 2020 to 10 February 2020. Similarly, the last 10 days (from 12 March 2020 to 21 March 2020) ahead forecast of cumulative confirmed cases from COVID-19 in the entire world are 210,000.

The well-known ARIMA model was used by Benvenuto et al. (29) to predict the trend of the spread and prevalence of novel coronavirus. Autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to estimate the parameters of the model. Based on the estimated values of the parameters, ARIMA (1,0,4) and ARIMA (1,0,3) were used to predict the prevalence and incidence of the COVID-19. The forecasted values based on the two ARIMA models of prevalence and incidence for the time window from 11 February 2020 to 02 February 2020 (2 days) are 45,151 and 2,418 with prediction intervals of (42,084 and 48,218) and (1,534 and 3,302), respectively.

An improved adaptive neuro-fuzzy inference system (ANFIS) based on an enhanced flower pollination algorithm (FPA) and slap swarm algorithm (SSA) was proposed by Al-Qaness et al. (30) to forecast the 10 days of cumulative confirmed cases from COVID-18 in China. The performance of the model has been increased by determining the parameters of both the ANFIS and FPASSA models. The efficiency of the proposed method in terms of RMSE, MAE, and MAPE is better than the other models. Based on the FPASSA-ANFS model, the estimated number of cumulative confirmed cases by 28 February 2020, in China is 99,453.

The well-known univariate time series ARIMA model was used to predict the cumulative confirmed cases, deaths, and recoveries from COVID-19 in Pakistan. Based on the investigational results of this study, ARIMA (0, 2, 1) (1, 0, 0) outperformed other time series models for predicting the next 10 days' cumulative confirmed cases. Similarly, ARIMA (0,2,1) was found to be the best candidate model for forecasting aggregate recoveries and deaths (31).

The problem of predicting the daily confirmed and daily deaths from COVID-19 has gained limited attention in the literature. Although some attempts have been made to address this issue, it is still a potential area to be investigated. Literature offers no clear methodology for the problem of predicting the daily confirmed and deaths from COVID-19. Here, we report a neglected aspect in previous studies, and an attempt has been made to address the issue with a more sophisticated and simple hybrid model. It can be observed from the graphical representation of the daily confirmed cases and deaths, as shown in **Figures 2**, **3**, which follow a nonlinear and nonstationary pattern that cannot be predicted easily by using any linear statistical or mathematical models.

To capitalize the strength of these models and address the issues and weaknesses of the abovementioned models, an attempt has been made to predict the daily confirmed cases and deaths from COVID-19 by suggesting a new hybrid EEMD-ETS model whose detailed description is outlined in the "Proposed hybrid EEMD-ETS model" section.

Since the daily confirmed cases and deaths from COVID-19 follow an irregular pattern, therefore, the traditional time series models might not enhance their nonlinear and stochastic characteristic and thus produce very unrealistic prediction results. This has been achieved primarily through the use of EEMD. The first step of this method is to decompose the nonlinear pattern of the data into dissimilar IMFs, and a single monotone residue component followed by the selected IMFs is then used to build the hybrid ETS model, which is then used for short-term prediction.

The novelty in this article is the development of a hybrid time series model which is based on the well-known idea of a divide-and-conquer algorithm that works recursively by breaking down the nonlinear COVID data into subgroups technically known as IMFs and then efficiently predicts COVID-19 in Italy, Germany, UK, and France.

The remaining article is organized in the following sections with techniques for future predictions in the "Prediction methods" section and the proposed hybrid EEMD-ETS model in the "Proposed hybrid EEMD-ETS model" section; experimental results on four COVID-19 datasets of Italy, France, Germany, and the UK are briefly explained in the "Experimental results" section, followed by discussion, and finally the conclusion is presented.

PREDICTION METHODS

In this section, we provided the details of the experimental procedures carried out in this study. Numerous research articles have shown that the time series forecasting model's emphasis on the past behavior of a random phenomenon best captures the underlying trends and patterns. The ideal model is then employed for the prediction of the future behavior of the underlying study variable. Over the past few years, there have been fabulous efforts carried out on the expansion of different time series models for forecasting the spread of contagions. In this article, we have suggested a hybrid technique that is based on EEMD and error trend seasonality (ETS) to predict the daily confirmed cases and deaths from COVID-19. A brief explanation of all the time series methods is outlined along with the proposed method in the following subsections.

Mean Method

In this method of forecasting, the mean value of all the historical time series is equal to the future forecast value. If we denote

the historical time series values by x_1, x_2, ..., x_t, then the future forecast value of *the k* period ahead is given by

$$\hat{x}_{t+k} = \underline{x} = (x_1 + x_2 + x_3 + \ldots + x_t)/t \quad (1)$$

Simple Exponential Smoothing

The simple exponential smoothing (SES) technique is one of the most common techniques of exponentially smoothing methods. Consider a time series x_1, x_2, ..., x_t with no seasonal or symmetric trend, the future forecasted value $\hat{x}_{t+k}$ is a weighted sum of the past values

$$\hat{x}_{t+k} = a_0 x_t + a_1 x_{t-1} + a_2 x_{t-2} + \ldots \quad (2)$$

where $\{a_i\}$ are weights in such a manner that more weights are given to the most recent values and fewer weights to the values that lie far away in the past. When the weights are increasing geometrically, the final equation for SES becomes

$$\hat{x}_{t+k} = \gamma x_t + \gamma (1-\gamma) x_{t-1} + \gamma (1-\gamma)^2 x_{t-2} + \ldots \quad (3)$$

Naïve Method

This method of forecasting works very efficiently for many economic and financial time series, especially when the time series follows random walks, that is why this method is sometimes known as the random walk forecasting method. In this method of point forecasting, the future forecast value is equal to the value of the last observation, i.e.,

$$\hat{x}_{t+k} = x_t \quad (4)$$

Theta Model

This theta model was proposed by Assimakopoulos and Nikolopoulos (32), where the basic idea of this forecasting method is altering the local curvature of the univariate time series through a coefficient known as "Theta" (θ) which is directly applied to the second difference of the time series. Therefore, a new series of time series known as Theta-lines are constructed and denoted as $L(\theta)$. Each of these Theta-lines is extrapolated individually and the forecasts are aggregated either equally weighted or through a weighed optimization procedure. Consider that the initial time series $Y = [Y_1, Y_2, Y_3, \ldots, Y_t]$ is decomposed into two Theta-lines, i.e., $L(\theta = 0)$ *and* $L(\theta = 2)$, then the algebraic equation for the model in its modified form is as follows:

$$Y_t = \frac{1}{2}\left(L_t(\theta = 0) + L_t(\theta = 2)\right), \quad \forall\, t = 1, 2, \ldots, n \quad (5)$$

TBATS Model

The trigonometric seasonality Box–Cox transformation ARIMA errors trend seasonal (TBATS) model developed by De Livera et al. (33) uses a combination of Fourier terms with an exponential smoothing state-space model and a Box–Cox transformation in an entirely automatic method. There is a slight difference between harmonic regression and the TBATS model, in the sense that the seasonal patterns are repeated without changing for the time in harmonic regression, while in the TBATS model, the seasonal components change slowly over time. The matrices for the TBATS model can be written as $= (1, \phi, a, \varphi, \theta)'$, $g = (\alpha, \beta, \gamma, 1, 0_{p-1}, 1, 0_{q-1})'$, and

$$F = \begin{bmatrix} 1 & \phi & 0_\tau & \alpha\phi & \alpha\theta \\ 0 & \phi & 0_\tau & \beta\phi & \beta\theta \\ 0'_\tau & 0'_\tau & A & B & C \\ 0 & 0 & 0_\tau & \varphi & \theta \\ 0'_{p-1} & 0'_{p-1} & O_{p-1,t} & I_{p-1,p} & O_{p-1,q} \\ 0 & 0 & 0_\tau & 0_p & 0_q \\ 0'_{q-1} & 0'_{q-1} & O_{q-1,\tau} & O_{q-1.p} & I_{q-1,q} \end{bmatrix} \quad (6)$$

Here, if all the components in the TBATS model are available, then these matrices are valid but if any of the components of the model is not available, then the corresponding term must be omitted from the matrices too.

The Holt-Winters Linear Trend Forecasting Procedure

This method of forecasting is the generalization of the SES technique by introducing two smoothing parameters α, γ for updating the local level (L_t) and trend (T_t) components of the time series. The values of these smoothing parameters generally fall in the range of (0, 1). The one forecast and two smoothing equations for the level and trend are given by

$$\hat{x}_{t+k} = L_t + kT_t \quad (7)$$

$$L_t = \alpha x_t + (1-\alpha)(L_{t-1} + T_{t-1}) \quad (8)$$

$$T_t = \gamma (L_t - L_{t-1}) + (1-\gamma) T_{t-1} \quad (9)$$

It can be seen from the level component given in equation 7 that L_t is a weighted average of the observations x_t and the one-step-ahead forecast given by $(L_{t-1} + T_{t-1})$. Similarly, the trend component given in equation 8 indicates that T_t is the weighted average of the estimated trend at time t based on $(L_t - L_{t-1})$ and L_{t-1}. The final k-step-ahead forecasted values are the linear combination of the last estimated level L_t and k times the last estimated trend values T_t.

Damped Trend Methods

The motivation behind this forecasting technique is the limitation of Holt's linear trend method that exhibits an endless trend component in the future horizon either increasing or decreasing those results in over-forecast, specifically for longer horizons. To overcome this drawback, Gardner and McKenzie (34) introduced a parameter that dampens the effect of the trend component in the future, and the values of this dampen parameter also lie in the range (0, 1). Mathematically, the holt-linear method is modified by incorporating the dampen parameter, i.e.,

$$\hat{x}_{t+k} = L_t + \left(\psi + \psi^2 + \psi^3 + \ldots + \psi^k\right) T_t \quad (10)$$

$$L_t = \alpha x_t + (1-\alpha)(L_{t-1} + \psi T_{t-1}) \quad (11)$$

$$T_t = \gamma (L_t - L_{t-1}) + (1-\gamma)\psi T_{t-1} \quad (12)$$

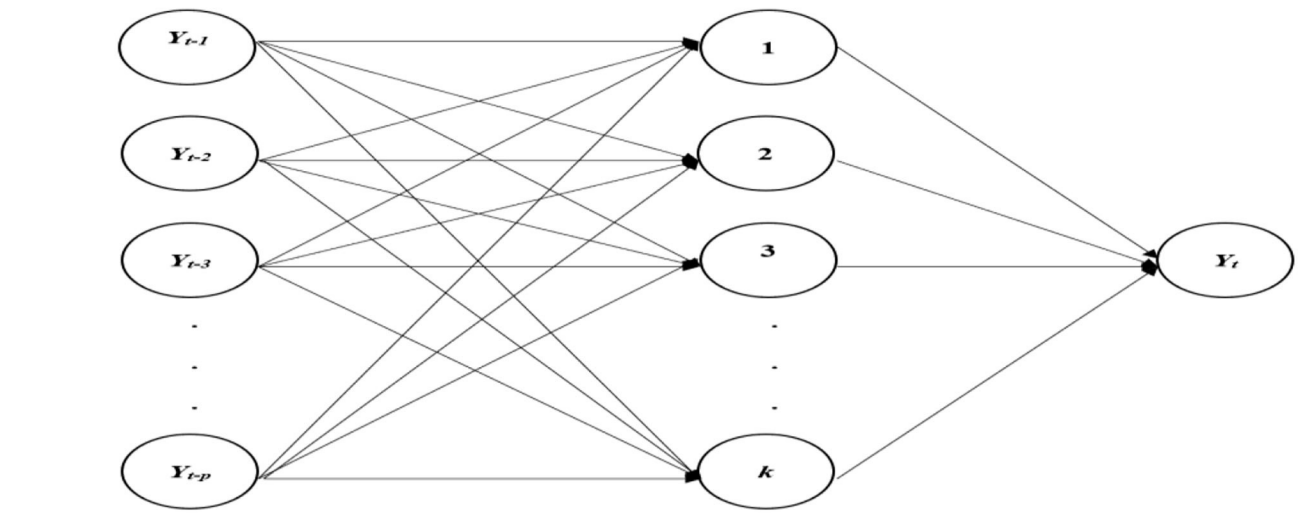

FIGURE 1 | NNAR model with *p* autoregressive terms as inputs and one hidden layer with *k* nodes.

ARIMA

This technique was first introduced by Box and Jenkins (35) and has been widely used for univariate time series forecasting. This method is completely data-driven, with the forecasted values of a variable depending upon the past or lagged values of the same variable. In terms of Y_t, the general forecasting equation is

$$Y_t = \beta + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \varepsilon_t - \varphi_1 \varepsilon_{t-1} - \varphi_2 \varepsilon_{t-2} - \ldots - \varphi_q \varepsilon_{t-q} \quad (13)$$

Here, the moving average parameters $\varphi's$ are described so that their signs are negative in the equation, following the convention presented by Box and Jenkins. Several researchers and software (i.e., the R) described them so that they have plus signs as an alternative. When real values are plugged into the equation, there is no doubt, but it is significant to distinguish which rules the software practices when interpreting the output. Often, the parameters are denoted by *AR (1), AR (2),…* and *MA (1), MA (2),* …. To recognize a suitable ARIMA model for Y_t, starting from the order of differencing *(d)* demanding to make the series stationary and eliminate the unstructured characteristics of seasonality, possibly in combination with a variance-stabilizing conversion, such as logging or deflating. If you end at this point and predict that the differenced series is constant, you have merely fitted a random walk or random trend model. However, the stationary series may still have auto-correlated errors, signifying that some values of *AR* terms $(p \geq 1)$ and/or some number *of MA* terms $(q \geq 1)$ are also desirable in the forecasting equation. The procedure of determining the values of p, d, and q that are excellent for a specified time series and plots of ACFs and partial autocorrelation functions (PACFs) will be used for this purpose.

Neural Network Autoregression

The ANNs are forecasting techniques that are founded on easy mathematical models of the brain. They permit compound nonlinear associations amid the response and predictor variables. A neural network is similar to a network of "neurons" which are ordered in layers. The predictors (or inputs) form the bottom layer, and the forecasts (or outputs) form the top layer. In the case of time series data, the lagged values can be used as inputs to the neural network and are known as the neural network autoregression (*NNAR*) model. In this study, we consider only the feed-forward neural network with one hidden layer denoted by *NNAR (p, k)*, meaning that there are p lagged inputs and k nodes in the hidden layer. A schematic diagram of the *NNAR* model is shown in **Figure 1**.

Long Short-Term Memory Model

Long short-term memory (LSTM) models belong to the artificial recurrent neural network (RNN) architecture that is widely used to handle sequence dependence in complex problem domains, namely, machine learning translations, speech recognition, handwriting recognition, and anomaly detection in network traffic of IDSs (intrusion detection systems). LSTM networks are compatible to classify, process, and make forecasts based on time series data with different activation functions, namely, sigmoid, hyperbolic, and hyperbolic tangents. LSTMs were established to solve the problem of vanishing gradients and exploding gradients that can be confronted by the traditional RNNs during the training phase (36). The most common LSTM network is comprised of a cell, an input gate, an output gate, and a forget gate. The cell recalls values over arbitrary time intervals, and the three gates control the flow of data into and out of the cell. The mathematical equations for the forward pass of an LSTM unit with forget

gate are

$$m_t = \vartheta_g \left(R_m y_t + P_m l_{t-1} + a_m\right) \quad (14)$$

$$n_t = \vartheta_g \left(R_n y_t + P_n l_{t-1} + a_n\right) \quad (15)$$

$$o_t = \vartheta_g \left(R_o y_t + P_o l_{t-1} + a_o\right) \quad (16)$$

$$\tilde{d}_t = \vartheta_g \left(R_d y_t + P_d l_{t-1} + a_d\right) \quad (17)$$

$$d_t = m_t \circ d_{t-1} + n_t \circ \tilde{d}_t) \quad (18)$$

$$l_t = o_t \circ \vartheta_l \left(c_t\right) \quad (19)$$

where the initial values of d_0 and l_0 are both equal to zero, and the operator $\circ$ is identified as a Hadamard product.

Error Trend and Seasonal Model

This method can forecast trends and seasonal components and is thus suitable for predicting the univariate time series. The ETS model is a special case of exponential smoothing models known as state-space models. There are different versions of these models which can be represented by the error, trend, and seasonality types, generally, a three-character string classifying method. The error component is denoted by the first letter ("A," "M," or "Z"); the trend type is represented by the second letter ("N," "A," "M," or "Z"); and the season type is represented by the third letter ("N," "A," "M," or "Z"). In all of these scenarios, "N" = none, "A" = additive, "M" = multiplicative, and "Z" = automatic selection. Therefore, for example, the SES with additive errors is denoted by "ANN"; similarly, the multiplicative Holt-Winters' method with multiplicative errors is "MAM" and so forth.

There are 30 models with different combinations of error, trend, and seasonality (37). **Supplementary Table 1** shows different combinations of these models. These models have fitted automatically to the data by using the method of maximum likelihood (ML) by optimizing the smoothing parameters and initial conditions with the help of a simple optimizer (38).

where N, M, A, A_d, and M_d denote None, Multiplicative, Additive, Damped Additive, and Damped multiplicative, respectively. Akaike's information criteria (AIC) and Bayesian information criteria (BIC) will be used for the selection of the best candidate ETS model. An ETS model with a minimum AIC value among the considered models will be chosen for application. The mathematical structure of AIC and BIC is given below (39, 40).

$$AIC = -2^* \ln \ln \left(l\right) + 2^* p \quad (20)$$

$$BIC = -2^* \ln \ln \left(l\right) + 2^* \ln \ln \left(n\right)^* p \quad (21)$$

Ensemble Empirical Mode Decomposition

The method of EMD uses the well-known Hilbert–Huang transform (HHT) technique to decompose the complex signal into dissimilar oscillatory components varying from low to high frequency and a single monotone residue (7). These oscillatory functions are technically known as IMFs. There are two basic conditions for each IMF: (i) the difference between the number of extrema and the number of zero-crossing will be one, and (ii) the upper and lower envelope will have zero mean. Given a signal *y(t)*, the algorithm of EMD can be used successfully to divide signals into their different components (11, 12, 41, 42). This method is robust, simple, and efficient that does not require any strong model assumptions. It is worthy to mention that some authors have used this method for the prediction of different complex and nonlinear time series datasets (43–45). The issue with this method relates to the problem of mode mixing which refers to the situation when an IMF resulting from EMD decomposition has components of different frequencies. Numerous efforts have been made on solving the problem of mode mixing and thus EEMD is one such alternative approach (8). This method has the flexibility to handle very complex signals without the mode mixing problem. In this technique, the white noise would be added to fill in the whole time–frequency space homogeneously, which can smooth an accepted separation of the frequency scales and diminish the existence of mode mixing. According to the properties of the EMD method, the procedure of EEMD can be described as follows:

Step 1: Add a random Gaussian white noise $n_i(t)$ to the original time series $y(t)$, the noise-added signal $y_i(t)$ is as follows:

$$y_i\left(t\right) = y\left(t\right) + n_i\left(t\right) \quad (22)$$

Step 2: Recognize all the local extrema (local maxima and minima) in the new signal $\{y_i\ (t)\}$.

Step 3: Find out the upper $\{U\ (t)\}$, and lower envelope $\{L\ (t)\}$ in the new white noise added signal $y_i(t)$.

Step 4: Join all the local extrema through the cubic spline interpolation technique to find out the mean of both the upper and lower envelope, i.e., $M\ (t)$:

$$Mean\left(t\right) = \frac{U\left(t\right) + L(t)}{2} \quad (23)$$

Step 5: The mean envelope calculated in step 4 will be subtracted from the actual signal to obtain the first component, i.e.,

$$k_1\left(t\right) = y\left(t\right) - Mean\left(t\right) \quad (24)$$

If $k_1\ (t)$ meets the two properties of the IMF defined above, then it should be well-thought-out as the first IMF; else, steps 1 to 5 will be repeated by considering $k_1\ (t)$ as a new-fangled signal.

Step 6: The first IMF obtained in step 5 will be deducted from the signal $y\ (t)$ to obtain $r_1\ (t)$, i.e.,

$$r_1\left(t\right) = y\left(t\right) - k_1\left(t\right) \ldots (25) \quad (25)$$

Step 7: In this step, $r_1\ (t)$ will be considered as a new signal and the sifting process of step 1 will be applied once again. The above process will continue until the last IMF is taken out from the signal. The overall trend of the signal will be a smooth monotonic residue obtained in the last step of EMD, and finally, the actual signal $y\ (t)$ can be decomposed as:

$$y\left(t\right) = \sum_{i=1}^{n} k_i\left(t\right) + r_n \ldots (26) \quad (26)$$

where r_n is the residue and $k_1(t)$, $k_2(t)$,…,$k_n(t)$ are different IMFs with different frequencies that vary from high to low. The final results of this decomposition are shown in **Supplementary Figures 1–4**.

It can be observed that the EEMD approach produces good quality results in terms of breaking the variations into their different components. In the first step, we decomposed the data into their different subparts varying from high- to low-frequency IMFs and a single monotone residual component. The results of this decomposition are presented and can be verified from **Supplementary Figures 1–4** given earlier. For all the four countries, seven (07) different IMFs are obtained for both the daily confirmed cases and daily deaths. After rigorous examination, it is revealed from these IMFs that there are two types of variation in the COVID-19 data, i.e., short term and long term. There are different reasons for short-term fluctuations that bring ups and downs in the daily confirmed cases and daily deaths, namely, imposing new restrictions, building emergency hospitals, and facilitating patients in intensive care units (ICU). These IMFs justify that any linear, mathematical, or statistical model will not produce good forecasting results unless they are used on the cumulative number of confirmed cases and the number of deaths.

PROPOSED HYBRID EEMD-ETS MODEL

The idea behind the proposed model is based on the well-famous divide-and-conquer algorithm that decomposes a given problem into multiple subproblems and their results are then combined efficiently. The proposed idea can be seen as a two-stage process, and the method of EEMD is implemented to decompose the nonlinear and nonstationary COVID-19 time series data into different IMFs in the first place and then the proposed method belongs to building the novel hybrid model in the second stage. The whole procedure is schematically shown in **Figure 2**, followed by a step-by-step implementation of the proposed hybrid model.

Step 1. The method of EEMD define above is used to decompose the actual COVID-19 daily confirmed cases and deaths data of all the four countries into different IMFs and residues.

Step 2. After decomposing the daily confirmed cases and deaths data into different IMFs and monotone residue in step 1, the proposed hybrid model is developed based on univariate time series ETS that belongs to the exponential family.

Step 3. In this step, the stationarity of each IMF is checked with the help of the augmented Dickey–Fuller (ADF) test (36). The ADF test is a well-known technique to test the null hypothesis that a unit root is present in the time series data. The alternative hypothesis is usually considered that the under-observation time series data are stationary. The results of this test are presented in **Supplementary Table 3**, **Tables 1–4**. After dividing IMFs into a non-overlapping sequence of stationary and nonstationary components, the overall mean of the stationary IMFs is subtracted from the actual data, to get the denoising signal, i.e.,

$$y_N(t) = x(t) - G.Mean[St(IMF(t)] \quad (27)$$

where $y_N(t)$ is the new denoised univariate time series data, $x(t)$ is the original data, and *G. Mean [SIMF(t)]* is the overall mean of the stationary IMFs.

Step 4. The univariate time series denoised signal is given as input to build the ETS model. The summary of each of these fitted ETS models is presented in **Table 4**, showing the corresponding values of smoothing parameters, the values of AIC and BIC, and the type of the best ETS model fitted.

Step 5. Once the ETS model is developed for the denoising data, the next step is to predict the future daily confirmed cases and deaths from COVID-19 for Italy, France, Germany, and the UK.

Step 6. Finally, the comparison is made between the predicted and hold-out datasets. Contrary to the traditional method of dividing the dataset into 80% training and 20% testing, the validity of this novel approach is demonstrated by using 259 observations out of 266 for model training, and the remaining 7 observations for checking its validity. Four statistical measures, i.e., root mean square error (RMSE), MAE, mean absolute percentage error (MAPE), and systematic mean absolute percentage error (sMAPE) (46) are used as a performance assessment criterion for the proposed model. The final results of these met1rics measures for the proposed and considered models are presented in **Supplementary Table 4**, **Tables 5–7** for all four countries.

EXPERIMENTAL RESULTS

Datasets

For this study, COVID-19 time series data on the number of daily confirmed cases, and the number of daily deaths were collected for four major European countries, i.e., Italy, United Kingdom (UK), Germany, and France from the website of the World Health Organization (WHO) during 23 February 2020, and 14 November 2020. A visual representation of the confirmed cases and deaths data for these countries is shown in **Figures 3**, **4**.

From **Figure 4**, it can be observed that the curves of the deaths flatten after attaining peak value, followed by a second spike that occurs in august after the relaxation in the restrictions across Europe. Keeping in view of the trajectory of COVID-19, these four countries implemented a second lockdown to stop the further spread of the virus and save the precious lives of their citizens.

The analysis of these figures suggests the nonlinear and nonlinear pattern of COVID-19 daily data for the number of confirmed cases and the number of deaths; therefore, it cannot

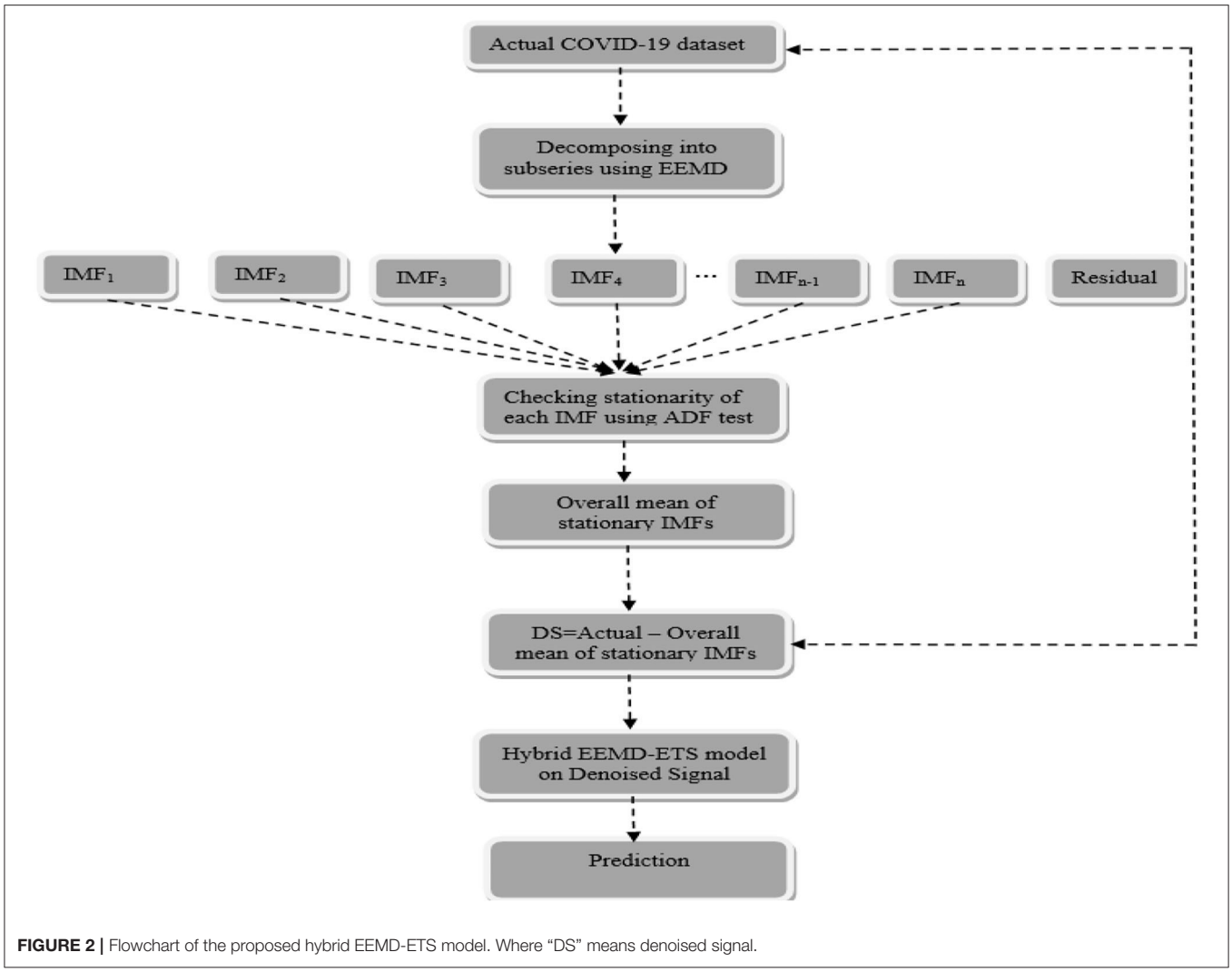

FIGURE 2 | Flowchart of the proposed hybrid EEMD-ETS model. Where "DS" means denoised signal.

be predicted more accurately with any linear time series or mathematical models. Thus, based on the nature of the data, a more robust technique is required to accurately predict the COVID-19 in these four countries.

An overall descriptive summary for the study variables is given in **Supplementary Tables 2, 3**, confirming that Italy, the UK, Germany, and France are the most affected countries by COVID-19 in Europe with more than 5 million cumulative confirmed cases and 151,380 cumulative deaths. Based on these statistics, it implies that Germany has taken all the protective measures issued by WHO to stop the spread of COVID-19 with only 2,908 average daily cases. Similarly, their health system efficiently managed the hospitalized patients which seems to be the only reasonable reason that is why the average of daily and cumulative deaths in Germany are slowed as compared to their neighboring countries. Similarly, the standard deviation (SD) of the daily confirmed cases and deaths is minimum for Germany, indicating that the data points tend to be very close to the average which showed their resilience against the contagious COVID-19.

Forecast Accuracy Criteria

The performance of the proposed approach and its resilience can be assessed by the following four statistical measures. The mathematical expressions of these four-performance metrics are given as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(A_t - P_{h,t}\right)^2} \tag{28}$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|A_t - P_{h,t}\right| \tag{29}$$

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{A_t - P_{t,h}}{A_t}\right| \ldots \tag{30}$$

$$sMAPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{\left|P_{t,h} - A_t\right|}{\left(|A_t| + \left|P_{t,h}\right|\right)^2} \tag{31}$$

where A_t and P_t denote actual and predicted values.

TABLE 1 | ADF test results along with the overall mean for Italy.

Component	ADF test value	*P*-value	Decision	Mean
$IMF1_{confirmedcases}$	1.9403	0.99	Non-stationary	Not required
$IMF1_{dailydeaths}$	−7.686	0.01	Stationary	−0.334
$IMF2_{confirmedcases}$	−10.868	0.01	Stationary	−11.384
$IMF2_{dailydeaths}$	−11.444	0.01	Stationary	0.102
$IMF3_{confirmedcases}$	−6.2013	0.01	Stationary	26.675
$IMF3_{dailydeaths}$	−3.922	0.0133	Stationary	−9.361
Overall Mean of daily confirmed cases	−9.49			
Overall Mean of daily deaths	−3.197			
$IMF4_{dailyconfirmed}$	−2.95	0.175	Non-stationary	Not required
$IMF4_{dailydeaths}$	−3.446	0.068	Non-stationary	Not required
$IMF5_{confirmed}$	−1.249	0.891	Non-stationary	Not required
$IMF5_{deaths}$	2.209	0.99	Non-stationary	Not required
$IMF6_{confirmed}$	−1.5308	0.773	Non-stationary	Not required
$IMF6_{deaths}$	2.84	0.99	Non-stationary	Not required
$IMF7_{confirmed}$	−0.049	0.99	Non-stationary	Not required
$IMF7_{deaths}$	1.578	0.99	Non-stationary	Not required

TABLE 2 | ADF test results along with the overall mean for France.

Component	ADF test value	*P*-value	Decision	Mean
$IMF1_{confirmedcases}$	−9.0304	0.01	Stationary	75.483
$IMF1_{dailydeaths}$	−7.842	0.01	Stationary	−6.631
$IMF2_{confirmedcases}$	−8.943	0.01	Stationary	−37.356
$IMF2_{dailydeaths}$	−7.006	0.01	Stationary	1.114
$IMF3_{confirmedcases}$	−5.846	0.01	Stationary	−35.881
$IMF3_{dailydeaths}$	−5.589	0.01	Stationary	5.192
$IMF4_{dailyconfirmed}$	−4.132	0.01	Stationary	115.706
$IMF4_{dailydeaths}$	−5.052	0.01	Stationary	−28.361
The overall mean of daily confirmed cases	29.488			
The overall mean of daily deaths	−7.171			
$IMF5_{confirmed}$	−0.59	0.9773	Not stationary	Not required
$IMF5_{deaths}$	1.067	0.99	Not stationary	Not required
$IMF6_{confirmed}$	−2.2627	0.4652	Not stationary	Not required
$IMF6_{deaths}$	2.243	0.99	Not stationary	Not required
$IMF7_{confirmed}$	0.0833	0.99	Not stationary	Not required
$IMF7_{deaths}$	−0.0373	0.99	Not stationary	Not required

TABLE 3 | ADF test results along with the overall mean for Germany.

Component	ADF test value	*P*-value	Decision	Mean
$IMF1_{confirmedcases}$	−6.904	0.01	Stationary	20.086
$IMF1_{dailydeaths}$	−7.607	0.01	Stationary	0.48
$IMF2_{confirmedcases}$	−7.353	0.01	Stationary	−8.189
$IMF2_{dailydeaths}$	−10.394	0.01	Stationary	0.2
$IMF3_{confirmedcases}$	−3.607	0.032	Stationary	12.276
$IMF3_{dailydeaths}$	−4.551	0.01	Stationary	0.619
$IMF4_{dailyconfirmed}$	−3.579	0.0356	stationary	−133.804
$IMF4_{dailydeaths}$	−3.396	0.055	Stationary	−7.759
Overall Mean of daily confirmed cases	−27.407			
Overall Mean of daily deaths	−1.614			
$IMF5_{confirmed}$	−0.965	0.9427	Not stationary	Not required
$IMF5_{deaths}$	1.078	0.99	Not stationary	Not required
$IMF6_{confirmed}$	−1.466	0.8005	Not stationary	Not required
$IMF6_{deaths}$	1.913	0.99	Not stationary	Not required
$IMF7_{confirmed}$	0.022	0.99	Not stationary	Not required
$IMF7_{deaths}$	−0.039	0.99	Not stationary	Not required

TABLE 4 | ADF test results along with the overall mean for UK.

Component	ADF test value	*P*-value	Decision	Mean
$IMF1_{confirmedcases}$	−5.529	0.01	Stationary	−29.176
$IMF1_{dailydeaths}$	−6.651	0.01	Stationary	−2.736
$IMF2_{confirmedcases}$	−8.915	0.01	Stationary	−2.778
$IMF2_{dailydeaths}$	−7.875	0.01	Stationary	0.385
$IMF3_{confirmedcases}$	−4.912	0.01	Stationary	−21.64
$IMF3_{dailydeaths}$	−4.523	0.01	Stationary	7.43
$IMF4_{dailyconfirmed}$	−3.813	0.018	Stationary	−15.715
$IMF4_{dailydeaths}$	−3.553	0.0381	Stationary	−24.06
The overall mean of daily confirmed cases	−17.33			
The overall mean of daily deaths	−4.745			
$IMF5_{confirmed}$	−1.457	0.8044	Not stationary	Not required
$IMF5_{deaths}$	1.51	0.99	Not stationary	Not required
$IMF6_{confirmed}$	−3.838	0.99	Not stationary	Not required
$IMF6_{deaths}$	1.1087	0.99	Not stationary	Not required
$IMF7_{confirmed}$	0.69	0.99	Not stationary	Not required
$IMF7_{deaths}$	0.1228	0.99	Not stationary	Not required

These metrics are commonly used techniques to evaluate the accuracy of different point forecasts of the competing models. The most popular and widespread is MAPE as it is very effective, easily understandable, and interpretable. It measures the prediction accuracy as a percentage and can be calculated as the average absolute percent error for each period minus actual values divided by actual values. Subsequently, RMSE, MAE, and sMAPE have also been extensively applied in the literature, although the interpretation of RMSE is more challenging to understand (47).

Analysis and Discussion

In this section, we discuss different time series model fittings, including the proposed hybrid model, and summarize the main findings of this study. All the COVID-19 data of the four countries were initially arranged into an excel sheet and

TABLE 5 | Performance of different models for 7 days prediction of Italy's daily confirmed cases.

Method	RMSE	MAE	MAPE	sMAPE
Mean	31998.353	31613.531	948.659	1.645
SES	5712.565	4365.085	11.54	0.124
Naïve	5712.774	4365.144	11.545	0.126
Theta	3526.853	2665.955	8.457	0.078
TBATS	4091.085	3568.494	9.636	0.107
HW	8790.664	7661.137	18.088	0.206
Damped	8200.137	6999.388	16.837	0.191
ETS	2552.256	2434.029	6.985	0.07
ARIMA	2711.679	2163.304	6.308	0.066
NNAR	4820.24	4019.41	11.667	0.112
LSTM	4874.481	3905.596	11.904	11.503
Hybrid EEMD-ETS	2404.163	1969.82	5.125	0.042

TABLE 6 | Performance comparison of different models for 7 days prediction of Italy's daily deaths.

Method	RMSE	MAE	MAPE	sMAPE
Mean	362.743	343.019	218.311	1.009
SES	139.556	121.623	28.576	0.252
Naïve	129.816	118.714	26.617	0.245
Theta	142.302	123.514	29.296	0.256
TBATS	87.509	75.007	14.687	0.158
HW	106.762	78.143	14.153	0.164
Damped	87.145	74.356	14.47	0.156
ETS	87.144	74.351	14.465	0.153
ARIMA	95.504	81.858	16.348	0.171
NNAR	115.198	105.808	22.691	0.218
LSTM	86.34	75.962	17.058	0.158
Hybrid EEMD-ETS	77.867	70.54	14.049	0.141

TABLE 7 | Performance comparison of different models for 7 days prediction of France's confirmed cases.

Method	RMSE	MAE	MAPE	sMAPE
Mean	37398.196	30671.534	487.759	1.317
SES	27427.455	26493.717	48.953	0.61
Naïve	31616.717	30863.438	51.238	0.675
Theta	27631.862	26714.523	49.135	0.614
TBATS	28348.173	27378.319	49.297	0.622
HW	31772.71	30910.284	52.713	0.677
Damped	27561.259	26395.92	49.811	0.609
ETS	31773.431	30910.951	52.71	0.677
ARIMA	26013.185	24794.244	47.679	0.581
NNAR	29952.856	27340.347	53.578	0.618
LSTM	27989.713	26956.326	49.987	0.618
Hybrid EEMD-ETS	23252.42	14253.33	40.654	0.353

then for further analysis, RStudio version 1.3.1093 and Python 3.3.6 with Jupyter Notebook were used. The data were first decomposed by successfully implementing the EEMD method into different IMFs, followed by finding out the stationary IMFs using the well-known ADF test. The results of which appear to tally with the authors' expectations that the high-frequency IMFs are mostly stationary and clustered around their mean.

From **Tables 1**–**4**, it can be seen that the ADF test results for which predefined value of $\alpha = 0.05$, the calculated p-value is less than the pre-specified alpha value that leads to the rejection of the null hypothesis that the given IMF is nonstationary, the nonstationary IMFs are not used to build our model, therefore their means are not required. The tabulated results of the ADF test confirm that the majority of the IMFs ranging from 1 to 4 for all the four countries, both for daily confirmed cases and daily deaths, are stationary. These are the most relevant findings and, perhaps, the most significant part of the composition of the proposed hybrid model based on EEMD and ETS approaches. The grand mean given in the tables presents the average short-term variations in the data. These short-term fluctuations are then subtracted from the original signal to get denoised COVID data as an ingredient for the ETS model followed by predictions. The values of smoothing parameters, AIC, BIC, and the type of models are presented in **Supplementary Table 4**.

At present, based on the minimum values of AIC and BIC, the best candidate ETS model is chosen for prediction, e.g., for Italy's daily confirmed cases, the best-reported model is ETS (M, Ad, M) which means that errors are multiplicative, and the trend in the data is damped additive with multiplicative seasonality. Similarly, to avoid repetition, the same description and interpretation can be made for other models as well.

Model Comparison

In this section, the proposed hybrid model is evaluated along with different selected time series models that demonstrate prediction results in the case of nonlinear and nonstationary COVID-19 data for the four selected countries. Here, we used a total of 11 methods; of these, 9 are conventional time series, one is a simple neural network with autoregressive terms (NNAR), and one is an RNN with LSTM), and the proposed hybrid method is based on EEMD and ETS models. We also checked the prediction performance of different potential hybrid models, namely, EMD-ARIMA, EMD-ETS, EEMD-ARIMA, and EEMD-ETS, of which the best hybrid model is chosen and the same is then compared with the competing models in terms of performance. To avoid confusion, we reported the best candidate model out of all potential hybrid models. The experimental results of the overall performance of these selected models are presented in terms of the following four measures, i.e., RMSE, MAE, MAPE, and sMAPE.

A key strength of this research lies in the fact that the prediction performance of our proposed model is equally efficient in all scenarios, i.e., for daily confirmed cases and daily deaths for all 4 countries. It can be verified easily from the investigational results presented in **Tables 5**–**12** that the four statistical measures of the suggested model are minimum. The values of RMSE for daily confirmed cases and deaths of Italy are **2404.13** and **77.86.** The second-best model based on the values of RMSE in this

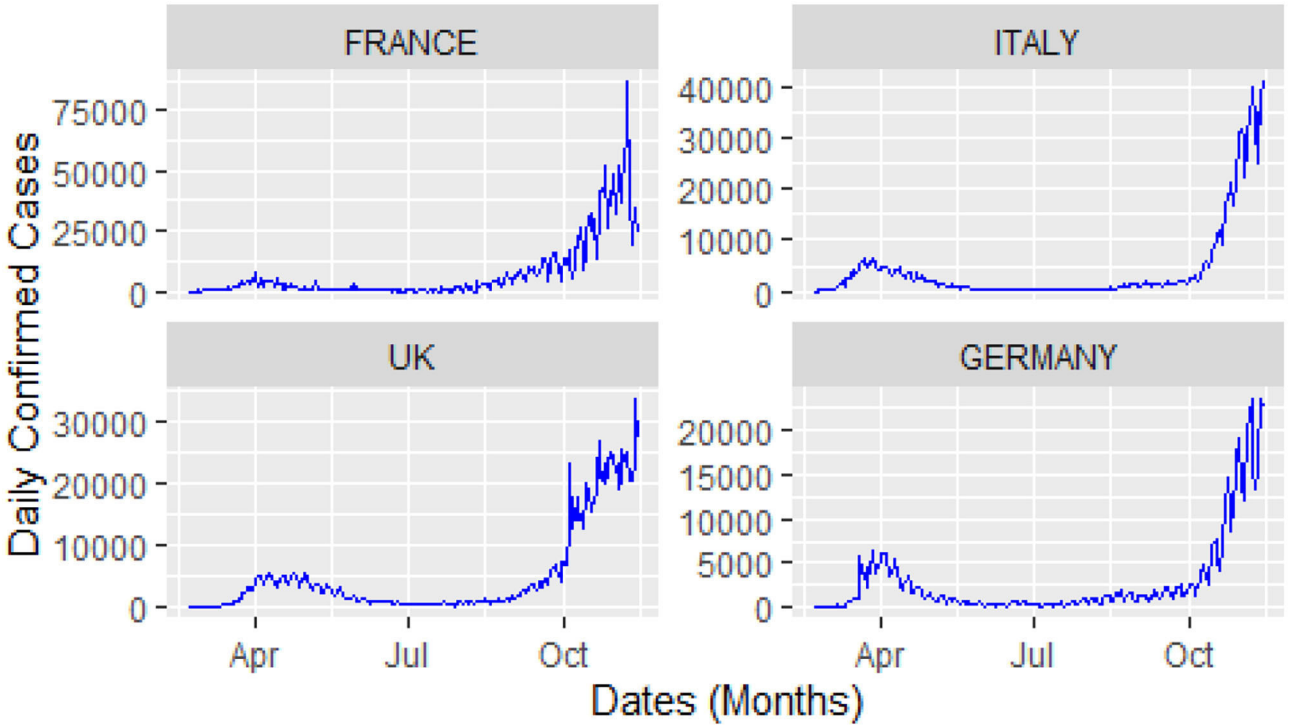

FIGURE 3 | Daily confirmed cases: time window from 23 February 2020 to 14 November 2020.

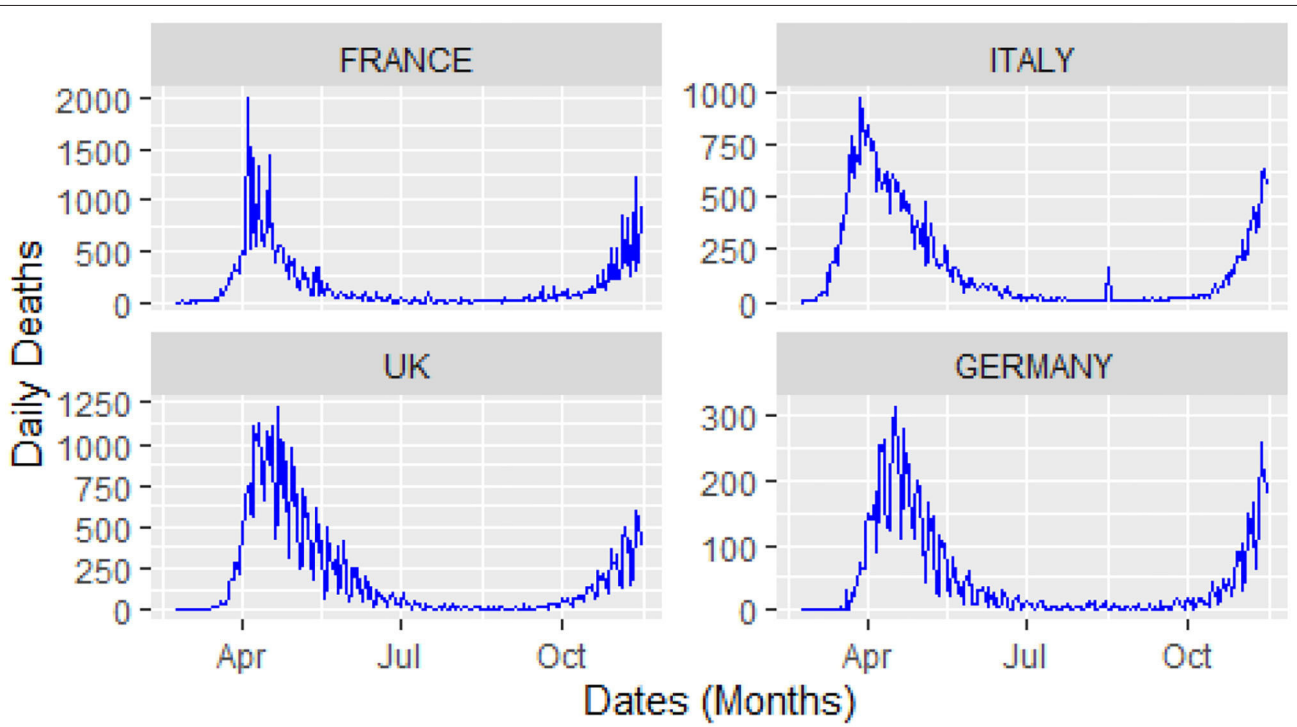

FIGURE 4 | Daily deaths: time window from 23 February 2020 to 14 November 2020.

TABLE 8 | Performance comparison of different models for 7 days prediction of France's daily deaths.

Method	RMSE	MAE	MAPE	sMAPE
Mean	539.424	420.586	275.178	0.996
SES	337.894	282.903	50.15	0.486
Naïve	422.35	395.571	47.832	0.626
Theta	281.609	214.945	40.914	0.366
TBATS	298.616	272.944	39.239	0.475
HW	359.555	328.171	44.652	0.55
Damped	328.118	298.227	46.379	0.51
ETS	299.161	265.928	42.413	0.461
ARIMA	288.136	244.529	41.183	0.426
NNAR	146.13	124.111	21.697	0.237
LSTM	335.04	310.608	46.969	0.528
Hybrid EEMD-ETS	102.733	82.378	14.101	0.146

TABLE 9 | Performance comparison of different models for 7 days prediction of Germany's confirmed cases.

Method	RMSE	MAE	MAPE	sMAPE
Mean	16654.38	16243.24	654.767	1.517
SES	5948.078	4715.722	20.153	0.24
Naïve	5948.227	4715.857	20.154	0.24
Theta	1385.131	1000.46	6.46	0.063
TBATS	1372.604	1008.138	6.015	0.529
HW	6933.016	6255.89	25.386	0.302
Damped	6652.431	5843.291	24.053	0.286
ETS	1772.325	1401.285	7.065	1.97
ARIMA	3043.065	2901.103	13.631	0.147
NNAR	2320.36	1904.506	10.035	0.109
LSTM	4593.045	3566.093	16.485	0.191
Hybrid EEMD-ETS	1298.967	935.492	5.348	0.053

competition is ETS with an RMSE value of **2552.25**, and the well-known ARIMA model stands in the third position with an RMSE of **2711.67.** Similarly, the values of MAE, MAPE, and sMAPE of our developed hybrid model are also minimum. Interestingly, the ARIMA model beats the ETS model in these metrics and stands in the second position in this forecast competition of COVID-19 for daily confirmed cases but failed to show good prediction results for daily deaths data (**Table 6**). In this scenario, the ETS model stands in the second position with minimum values of MAE, MAPE, and sMAPE after the proposed hybrid model.

The experimental results for France's daily deaths and daily confirmed cases presented in **Tables 7**, **8** show that our model outperformed other models in this forecast competition, while the well-known ARIMA model's performance is much better than his strong rival, the ETS model, and stands with the second position with minimum values of RMSE, MAE, and sMAPE.

Investigational results for Germany and the UK are shown in **Tables 9**–**12**. It can be verified from the values of four

TABLE 10 | Performance comparison of different models for 7 days prediction of Germany's daily deaths.

Method	RMSE	MAE	MAPE	sMAPE
Mean	140.659	121.227	279.69	1.024
SES	79.27	72.855	56.041	0.505
Naïve	79.271	72.857	56.043	0.505
Theta	51.394	40.253	27.565	0.228
TBATS	38.225	30.126	20.943	0.23
HW	59.52	48.717	28.418	0.349
Damped	58.292	49.025	30.143	0.35
ETS	37.547	32.468	22.259	0.24
ARIMA	47.14	37.354	26.018	0.258
NNAR	91.139	1117.111	20.697	0.302
LSTM	61.177	51.285	35.68	0.362
Hybrid EEMD-ETS	17.604	15.01	10.258	0.105

TABLE 11 | Performance comparison of different models for 7 days prediction of UK's daily cases.

Method	RMSE	MAE	MAPE	sMAPE
Mean	20472.959	20003.733	451.902	1.373
SES	4459.872	3418.096	14.555	0.136
Naïve	4505.504	3390.143	14.558	0.134
Theta	4165.609	3253.262	14.716	0.133
TBATS	4355.468	3380.606	14.399	0.134
HW	4084.515	3444.378	13.839	0.137
Damped	4184.243	3437.578	14.117	0.137
ETS	3954.365	3333.344	13.418	0.133
ARIMA	4288.603	3226.114	13.809	0.127
NNAR	4686.289	3288.208	14.549	0.13
LSTM	4515.138	3395.429	14.597	0.135
Hybrid EEMD-ETS	4076.516	3130.997	12.573	0.123

TABLE 12 | Performance comparison of different models for 7 days prediction of UK's daily deaths.

Method	RMSE	MAE	MAPE	sMAPE
Mean	271.024	225.884	120.689	0.663
SES	167.729	148.916	41.115	0.409
Naïve	169.672	152.98	42.816	0.418
Theta	119.826	88.672	26.12	0.216
TBATS	114.393	109.527	29.814	0.308
HW	154.058	132.409	29.813	0.364
Damped	151.3	137.454	33.513	0.379
ETS	75.049	68.648	19.067	0.201
ARIMA	73.45	61.904	17.049	0.189
NNAR	95.206	105.765	41.987	0.487
LSTM	154.763	141.574	36.855	0.391
Hybrid EEMD-ETS	70.954	60.976	15.711	0.118

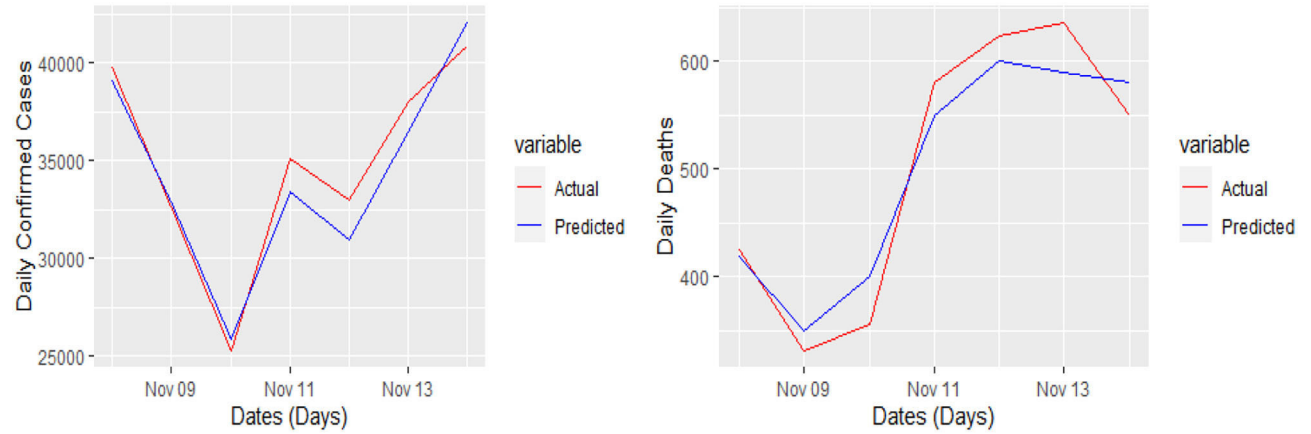

FIGURE 5 | Actual and predicted 7 days daily confirmed cases and deaths for Italy from COVID-19.

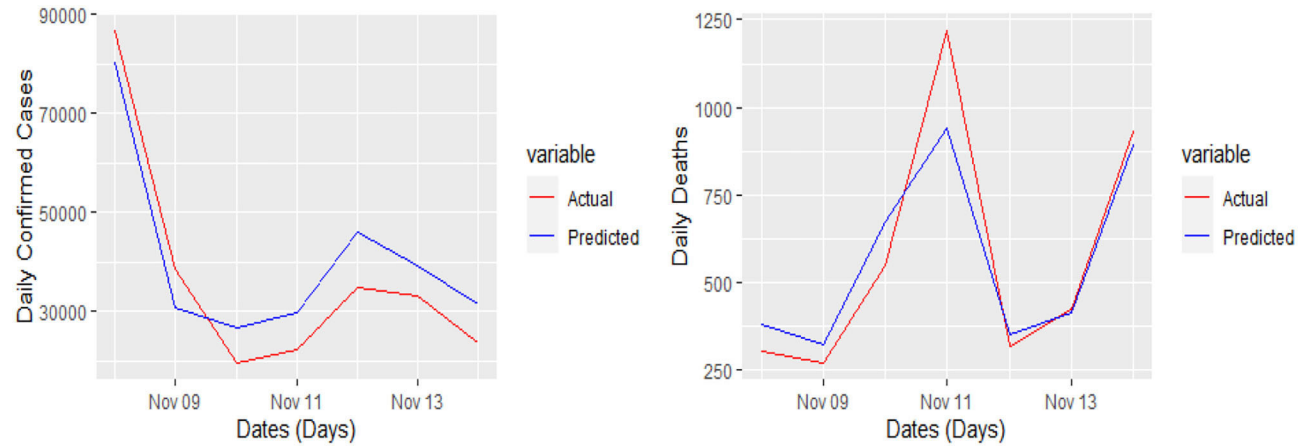

FIGURE 6 | Actual and predicted 7 days daily confirmed cases and daily deaths in France from COVID-19.

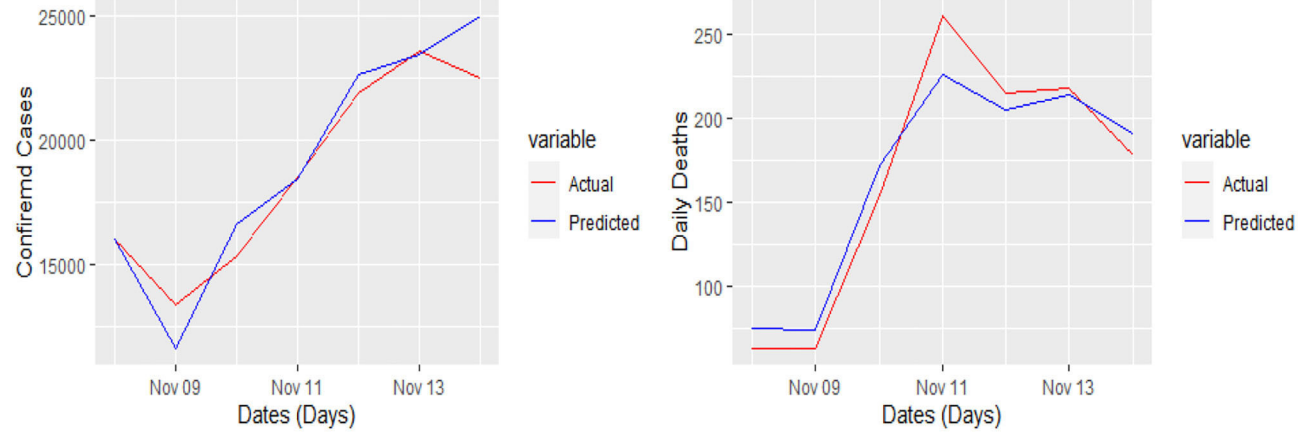

FIGURE 7 | Actual and predicted 7 days daily confirmed cases and daily deaths from COVID-19 in Germany.

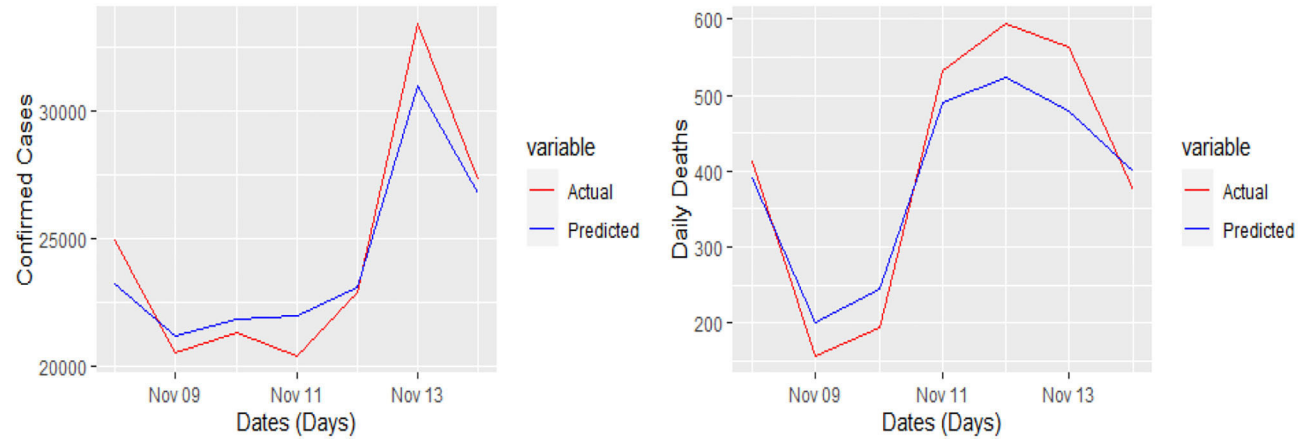

FIGURE 8 | Actual and predicted 7 days daily confirmed cases and daily deaths from COVID-19 in the UK.

statistical metrics, i.e., RMSE, MAE, MAPE, and sMAPE, that the prediction performance of our suggested model is better than the other conventional and machine learning methods. The ETS model again outperformed the classical ARIMA model and holds the second position for Germany and the UK.

The 7 days prediction was made by implementing our proposed model. To save space, we are not reporting these values here; a snapshot will better reflect the scope of our study. To make the prediction clear and understandable, we presented the actual and predicted values schematically through **Figures 5–8** for each country and each case. In all these cases, the actual and predicted daily confirmed cases and daily deaths are denoted by solid red and blue lines, respectively.

Yet, the actual and predicted values are far from each other but the direction accuracy of our prediction is more than 90%, which can be verified through **Figures 5–8**; such a tremendous direction accuracy will help the governments for better policies to stop the spread of the pandemic.

CONCLUSION

Prediction of the pandemics is always interesting, and there are numerous areas of research for data practitioners. Accurate prediction of pandemics is of great importance as it will help the governments to implement their resources in a better manner to stop the spread and save the precious lives of their citizens. The main conclusion of this study is drawn together and presented in this section. In most of the previous studies, the researchers used a single mathematical or statistical model to predict the accurate trajectory of the COVID-19 and, therefore, criticized for its poor prediction performance. The key objective of this research work is to propose a novel method to predict the contagious COVID-19 daily confirmed cases and deaths in four major European countries, i.e., Italy, France, Germany, and the UK. A key strength of this research lies in the fact that we proposed a hybrid method that is based on EEMD and univariate time series ETS model. Thus, the suggested technique is very appropriate for prediction with nonlinear and nonstationary data. Our proposed model is not an ensemble model as we did not utilize all the subcomponents after decomposing the COVID-19 data into different IMFs and single monotone residual by implementing the method of EEMD, we used only stationary IMFs to build our model. After successfully implementing the model, we used it for short-term forecasting of only 7 days. A comparison is made with other conventional univariate time series, NNAR and LSTM models. Based on the investigational results of the four statistical metrics, i.e., RMSE, MAE, MAPE, and sMAPE, the proposed model outperformed the other models, indicating that it is a promising tool for COVID-19 prediction. Surprisingly, the univariate single ETS and ARIMA model stands second in this competition and outperformed the NNAR and LSTM model, while we were expecting that the deep neural network LSTM model will perform better than the traditional univariate time series models except the suggested one.

In the future, we are looking to use our proposed algorithm for other countries' COVID-19 data by using different variables, namely, daily recoveries, daily hospitalized patients, and spread rate as well as to check its performance on other univariate time series datasets, namely, stock returns, exchange rates, wind speed, temperature, rainfall, earthquakes, tourist arrival, and crude oil. In short, we are planning to test the accuracy of our proposed model on any nonlinear and nonstationary univariate time series data.

Our study has some drawbacks that require additional investigation. First, as the data are very limited, therefore, the performance of the model will be checked by using it on a longer series and long-term forecasting.

To end with, in this research article, we proposed a hybrid EEMD-ETS model to predict the daily confirmed cases and daily deaths from the current pandemic of COVID-19 using Italy, France, Germany, and UK datasets.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: The data is already included in the article.

AUTHOR CONTRIBUTIONS

DK, AAl, and AAf: conceptualization. DK, HA, and AAf: formal analysis and methodology. HA: funding acquisition. DK, MA, NI, and UK: investigation, resources, and writing–original draft. AAf: project administration. DK and UK: software. DK: supervision. DK, HA, and AAl: validation. MA and NI: visualization. HA, AAl, and AAf: writing–review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by Taif University Researchers Supporting Project number (TURSP-2020/279), Taif University, Taif, Saudi Arabia.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2022.922795/full#supplementary-material

REFERENCES

1. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *N Engl J Med*. (2020) 382:1199–207. doi: 10.1056/NEJMoa2001316
2. World Health Organization. Available online at: https://covid19.who.int/table.
3. Lai D. Monitoring the SARS epidemic in China: a time series analysis. *J Data Sci.* (2005) 3:279–93. doi: 10.6339/JDS.2005.03(3)0.229
4. Wagenaar BH, Augusto O, Beste J, Toomay SJ, Wickett E, Dunbar N, et al. The 2014–2015 Ebola virus disease outbreak and primary healthcare delivery in Liberia: time-series analyses for 2010 2016. *PLoS Med.* (2018) 15:e1002508. doi: 10.1371/journal.pmed.1002508
5. Earnest A, Tan SB, Wilder-Smith A, Machin D. Comparing statistical models to predict dengue fever notifications. *Comput Math Methods Med.* (2012) 2012:758674. doi: 10.1155/2012/758674
6. Kane MJ, Price N, Scotch M, Rabinowitz P. Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1outbreaks. *BMC Bioinformatics.* (2014) 15:276. doi: 10.1186/1471-2105-15-276
7. Al-Babtain AA, Gemeay AM, Afify AZ. Estimation methods for the discrete poisson-lindley and discrete lindley distributions with actuarial measures and applications in medicine. *J King Saud Univer Sci.* (2021) 33:101224. doi: 10.1016/j.jksus.2020.10.021
8. Liu X, Ahmad Z, Gemeay AM, Abdulrahman AT, Hafez EH, Khalil N, et al. Modeling the survival times of the COVID-19 patients with a new statistical model: a case study from China. *PLoS ONE.* (2021) 16:e0254999. doi: 10.1371/journal.pone.0254999
9. Alzeley O, Almetwally EM, Gemeay AM, Alshanbari HM, Hafez EH, Abu-Moussa MH, et al. Statistical inference under censored data for the new exponential-X fréchet distribution: simulation and application to leukemia data. *Comput Intell Neurosci.* (2021) 2021:2167670. doi: 10.1155/2021/2167670
10. Teamah AEA, Elbanna AA, Gemeay AM. Fréchet-Weibull mixture distribution: properties and applications. *Appl Math Sci.* (2020) 14:75–86. doi: 10.12988/ams.2020.912165
11. Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis, *Proc R Soc Lond Math Phys Eng Sci A*. (1998) 454:903–95. doi: 10.1098/rspa.1998.0193
12. Wu Z, Huang NE. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv Adapt Data Anal.* (2009) 1:1–41. doi: 10.1142/S1793536909000047
13. Huang NE, Shen Z, Long SR. A new view of nonlinear water waves: the Hilbert spectrum, Annu. *Rev Fluid Mech.* (1999) 31:417–57. doi: 10.1146/annurev.fluid.31.1.417
14. Vincent, S-Hu LJ HT, Hou Z. Damage detection using empirical mode decomposition method and a comparison with wavelet analysis. In: *Proceedings of the Second International Workshop on Structural Health Monitoring*. Stanford, CA (1999). p. 891–900.
15. Yu DJ, Cheng JS, Yang Y. Application of EMD method and Hilbert spectrum to the fault diagnosis of roller bearings, Mech. *Syst. Signal Process.* (2005) 19:259–270. doi: 10.1016/S0888-3270(03)00099-2
16. Junsheng C, Dejie Y, Yu Y. The application of energy operator demodulation approach based on EMD in machinery fault diagnosis. *Mech Syst Signal Process.* (2007) 21:668–77. doi: 10.1016/j.ymssp.2005.10.005
17. Zhou F, Zhou HM, Yang Z, Yang L. EMD2FNN: a strategy combining empirical mode decomposition and factorization machine based neural network for stock market trend prediction. *Expert Syst Appl.* (2019) 115:136–51. doi: 10.1016/j.eswa.2018.07.065
18. Lin CS, Chiu SH, Lin TY. Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting. *Econ Model.* (2012) 29:2583–90. doi: 10.1016/j.econmod.2012.07.018
19. Yu L, Wang S, Lai KK. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. *Energy Economics.* (2008) 30:2623–35. doi: 10.1016/j.eneco.2008.05.003
20. Al-Babtain AA, Elbatal I, Al-Mofleh H, Gemeay AM, Afify AZ, Sarg AM, et al. The flexible burr XG family: properties, inference, and applications in engineering science. *Symmetry.* (2021) 13:474. doi: 10.3390/sym13030474
21. Zhao S, Musa SS, Lin Q, Ran J, Yang G, Wang W, et al. Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: a data driven modelling analysis of the early outbreak. *J Clin Med.* (2020) 9:388. doi: 10.3390/jcm9020388
22. Tang B, Wang X, Li Q, Bragazzi NL, Tang S, Xiao Y, et al. Estimation of the transmission risk of the 2019-nCoV and Its implication for public health interventions. *J Clin Med.* (2020) 9:462. doi: 10.3390/jcm9020462
23. Ture M, Kurt I. Comparison of four different time series methods to forecast hepatitis A virus infection. *Expert Syst. Appl.* (2006) 31:41–6. doi: 10.1016/j.eswa.2005.09.002
24. Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy, and France. *Chaos Solitons Fractals.* (2020) 134:109761. doi: 10.1016/j.chaos.2020.109761
25. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. *Infectious Dis Model.* (2020) 5:256–263. doi: 10.1016/j.idm.2020.02.002
26. Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? *Lancet.* (2020) 395:P1225–8. doi: 10.1016/S0140-6736(20)30627-9
27. Anastassopoulou C, Russo L, Tsakris A, Siettos Data-based analysis C. modeling, and forecasting of the COVID-19 outbreak. *PLoS ONE.* (2020) 15:0230405. doi: 10.1371/journal.pone.0230405
28. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. *PLoS ONE.* (2020) 15:e0231236. doi: 10.1371/journal.pone.0231236
29. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data Brief.* (2020) 2020:105340. doi: 10.1016/j.dib.2020.105340

30. Al-Qaness MA, Ewees AA, Fan H, Abd El Aziz M. Optimization method for forecasting confirmed cases of COVID-19 in China. *J Clin Med.* (2020) 9:674. doi: 10.3390/jcm9030674
31. Ali M, Khan DM, Aamir M, Khalil U, Khan Z. Forecasting COVID-19 in Pakistan. *PLoS ONE.* (2020) 15:e0242762. doi: 10.1371/journal.pone.0242762
32. Assimakopoulos V, Nikolopoulos K. The theta model: a decomposition approach to forecasting. *Int J Forecast.* (2000) 16:521–30. doi: 10.1016/S0169-2070(00)00066-2
33. De Livera AM, Hyndman RJ, Snyder RD. Forecasting time series with complex seasonal Patterns using exponential smoothing. *J Am Stat Assoc.* (2011) 106:1513–27. doi: 10.1198/jasa.2011.tm09771
34. Gardner Jr ES, McKenzie ED. Forecasting trends in time series. *Manage Sci.* (1985) 31:1237–46. doi: 10.1287/mnsc.31.10.1237
35. Box G, Jenkings G. *Time Series Analysis*: San Francisco, CA: Forecasting and Control, Holden Day (1970).
36. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. In: *9th International Conference on Artificial Neural Networks: ICANN '99*. Edinburgh, UK (1999).
37. Makridakis S. Accuracy measures: theoretical and practical concerns. *Int J Forecast.* (1993) 9:527–9. doi: 10.1016/0169-2070(93)90079-3
38. Hyndman RJ, Khandakar Y. *Automatic Time Series for Forecasting: The Forecast Package for (No. 6/07)*. Clayton VIC: Monash University, Department of Econometrics and Business Statistics (2007).
39. Stone M. Comments on model selection criteria of Akaike and Schwarz. *J R Stat Soc B.* (1979) 41:276–8. doi: 10.1111/j.2517-6161.1979.tb01084.x
40. Akaike HA. new look at the statistical model identification. *IEEE Trans Automat Contr.* (1974) 19:716–23. doi: 10.1109/TAC.1974.1100705
41. Huang NE. Review of empirical mode decomposition. In: *Proceedings of SPIE.* Udine, Italy. (2001). 71–80.
42. Wu, Hu CK MC. Empirical mode decomposition and synchrogram approach to cardiorespiratory synchronization. *Phys Rev E.* (2006) 73:051917. doi: 10.1103/PhysRevE.73.051917
43. Kang A, Tan Q, Yuan X, Lei X, Yuan Y. Short-term wind speed prediction using EEMD LSSVM model. *Adv Meteorol.* (2017) 2017. doi: 10.1155/2017/6856139
44. Wang WC, Chau KW, Xu DM, Chen XY. Improving forecasting accuracy of annual runoff timeseries using ARIMA based on EEMD decomposition. *Water Resour Manag.* (2015) 29:2655–2. doi: 10.1007/s11269-015-0962-6
45. Yu L, Dai W, Tang L. A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting. *Eng Appl Artif Intell.* (2016) 47:110–21. doi: 10.1016/j.engappai.2015.04.016
46. Dickey D, Fuller WA. Distribution of the estimators for time series regressions with a unit root. *J Am Stat Assoc.* (1979) 74:427–31. doi: 10.1080/01621459.1979.10482531
47. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice.* Melbourne, Australia: OTexts (2018).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Khan, Ali, Iqbal, Khalil, Aljohani, Alharthi and Afify. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.