State Estimation of the Spread of COVID-19 in Saudi Arabia using Extended Kalman Filter

Lamia Alyami
Department of Mathematics, College of Engineering,
Mathematics and Physical Sciences, University of Exeter,
Penryn Campus TR10 9FE, United Kingdom.
Email: la424@exeter.ac.uk

Saptarshi Das, *Member, IEEE*
Department of Mathematics, College of Engineering,
Mathematics and Physical Sciences, University of Exeter,
Penryn Campus TR10 9FE, United Kingdom.
Email: S.Das3@exeter.ac.uk, saptarshi.das@ieee.org

***Abstract*—COVID-19 has caused global concern as the World Health Organization (WHO) considered it a global pandemic that has affected all countries to different extent. Numerous studies have examined the behaviour of the pandemic using a wide variety of mathematical models. In this paper, we consider the nonlinear compartmental epidemiological dynamical system model in the Susceptible-Exposed-Infected-Quarantined-Recovered-Deceased (SEIQRD) form based on the recursive estimator known as the extended Kalman filter (EKF) to predict the evolution of the COVID-19 pandemic in Saudi Arabia. We adopt the nested sampling algorithm for parameter estimation and uncertainty quantification of the SEIQRD model parameters using real data. Our simulation results show that the EKF can not only predict the evolution of the directly measured variables i.e. the total death (*D*) and active case (*I*) but can also be useful in the estimation of the unmeasurable state variables and help predicting their future trends.**

***Index Terms*—Extended Kalman Filter (EKF), SEIQRD model**

I. Introduction

Coronaviruses SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) or the COVID-19 are a group of viruses that has become a global pandemic recognised by the WHO, affecting every country. The COVID-19 virus has high transmissibility which has caused infections and increased the burden on the public health, especially for older people. The WHO declared the COVID-19 as a global pandemic in March 2020 [1] which started a rapid increase in research in epidemiological modeling using different countries' data. Now, with more mutated variants of this virus, it is crucial to investigate and predict the long-term behaviour of such infectious disease models from an early stage until the end of the pandemic. In March 2022 there were more than 400 million confirmed cases of the coronavirus and the death cases have exceeded 6 million. The fundamental epidemiological model called SIR *(Susceptible-Infected-Recovered)* introduced by [2] has been utilized widely to predict the COVID-19 outbreak. SIR and its modifications such as SEIR *(Susceptible-Exposed-Infected-Recovered)* etc., have been discuss in details in [3]. Developing the SEI model within Bayesian framework and estimate the posterior probability distributions for parameters of interest using Markov chain Monte Carlo (MCMC) method was reported in [4]. The SEIR model in [5] demonstrated a consistent prediction between the model outputs and real COVID-19 data in Saudi Arabia. Later in [6] the SEIQR *(Susceptible-Exposed-Infected-Quarantine-Recovered)* model has been studied for stability analysis of the model based on Saudi Arabia infection daily data. The SEIQRD *(Susceptible-Exposed-Infected-Quarantine-Recovered-Dead)* model was used in [7] for risk management to forecast the spread of COVID-19 pandemic in different countries.

As opposed to these existing works, the aim of this paper is to construct a new epidemiological model SEIQRD with reinfection to explain the long-term COVID-19 spread in Saudi Arabia. This will then enhance the prediction using Kalman filtering algorithms and help understand the behaviour of the underlying dynamic process and track hidden or unmeasurable variables of the compartmental model. With this aim, we have investigated the utility of the KF family of algorithms in COVID-19 spread. The KF algorithm is a recursive technique to generate estimates of the state variables based on measured time-series data of a fewer variables (e.g. active case and total death) often corrupted with noise and bias. The KF algorithm is well-known as an optimal recursive solution of the discrete-time linear observer or state estimation problem [8]. The optimality of the KF is in the sense of minimizing the mean squared error (MMSE). The KF algorithm operates in two steps: 1) predicting the current state estimates from the previous estimates; 2) updating the estimates of the filter using the error covariance matrices with the feedback mechanism. Hence, the KF is referred as a predictor-corrector algorithm.

The drawback of the classical KF algorithm is strictly applicable to the linear systems and Gaussian noise. For this reason KF has been extended to other different versions. One of its variant solves the model nonlinearity problem, known as the Extended Kalman Filter (EKF). The EKF is an approximation filter for the nonlinear systems based on first-order Taylor series expansion, evaluated at each time step around the current state. More information on EKF can be found in [9], [10]. The EKF can be applied in tracking the epidemiological processes but there has been limited amount of literature available, addressing the use of EKF in tracking the spread of other types of infectious diseases e.g. [11], [12]. For using the EKF to estimate the COVID-19 spread, there are only few literature addressing the COVID-19 pandemic. In [13], the

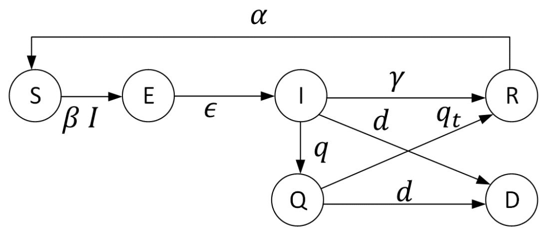

Fig. 1: The proposed SEIQRD model with reinfection.

authors used the maximum likelihood estimation (MLE) in the EKF to predict the COVID-19 transmission in China and USA. Also, [14] proposed the use of EKF to simulate stochastic and deterministic models with greater accuracy for the prediction of COVID-19 behaviour. In [15] the authors have implemented the EKF to predict the evolution of the COVID-19 pandemic, within a short time span. Since the rapid outbreak of the COVID-19, the accuracies of different dynamic models play an important role in prediction and decision-making. Here, the COVID-19 pandemic data has been used to fit a newly proposed model of SEIQRD form with reinfection term along with quantifying parametric uncertainties. We have considered a reinfection term in the model since the reinfections cases are reported in different sources e.g. [16]. Based on this model, the EKF has been used to analyze the COVID-19 behaviour over long-term. The simulation of the proposed method is applied for the COVID-19 data outbreak in Saudi Arabia from 15 February 2020 to 17 March 2022. Moreover, the real data has been compared with the EKF predictions on the measured states beside estimating the trends of the unmeasurable states of this SEIQRD compartmental model.

II. Description of the Proposed SEIQRD Model

The SEIQRD model divides the population into six classes: susceptible $S(t)$, exposed $E(t)$, infected $I(t)$, quarantined $Q(t)$, recovered $R(t)$ and dead $D(t)$. The compartments were modelled using the system illustrated in Fig. 1 as the nonlinear differential equation or nonlinear state space model as follows:

$$\begin{aligned} \frac{dS}{dt} &= -\beta IS + \alpha R, \\ \frac{dE}{dt} &= \beta IS - \epsilon E, \\ \frac{dI}{dt} &= \epsilon E - \gamma I - qI - dI, \\ \frac{dQ}{dt} &= qI - q_t Q - dQ, \\ \frac{dR}{dt} &= \gamma I + q_t Q - \alpha R, \\ \frac{dD}{dt} &= dI + dQ. \end{aligned} \tag{1}$$

The model parameters $\{\beta,\ \gamma,\ \epsilon,\ q,\ q_t,\ \alpha,\ d\}$ are non-negative and defined as the infection rate, recovery rate, incubation rate, quarantine rate, quarantine period, reinfection rate, and death rate respectively. Since the length of the protective immunity is unknown, we consider the possibility of reinfection after recovery where a fraction (α) of the recovered population returns to the susceptible compartment and when α = 0 our proposed model coincides with the SEIQRD model proposed in [17]. It is a closed compartmental model (2) defined as:

$$S + E + I + Q + R + D = N, \tag{2}$$

defining N as the total size of the population of a country under study. The basic reproduction number R_0 for the proposed model can be defined based on the next generation matrix proposed in [18] as:

$$R_0 = \frac{N\beta}{d + \gamma + q}. \tag{3}$$

III. Methodology for Model Parameter and Uncertainty Estimation

The official epidemic data is frequently published on a daily or weekly basis. The reported measurements are usually in discrete time domain (cases per day) whereas the epidemiological model is in continuous time. Therefore, the model under local linearization needs to be discretized to match with the dataset collected on a daily basis using the hybrid extended Kalman filter (discrete-continuous EKF). The first step of this is estimating the parameters of the SEIQRD model using a Bayesian uncertainty quantification method called the nested sampling algorithm that draws samples from the posterior distribution of the unknown dynamic model parameters as a by-product while calculating the Bayesian evidence or the marginal likelihood [19]. Using the mean estimate of the posterior distribution of the unknown model parameters, we carry out the state estimation using the EKF approach. We consider the prediction of COVID-19 spread using the locally linearized hybrid EKF using the mean posterior of the new SEIQRD model parameters to estimate the unmeasurable states with a given initial state vector X_0 and covariance matrix P_0.

IV. Extended Kalman Filter Applied to the SEIQRD Model for State Estimation

The system state vector X for the SEIQRD model with reinfection (1) is defined as:

$$X = \begin{bmatrix} S & E & I & Q & R & D \end{bmatrix}^T. \tag{4}$$

Starting with the non-negative initial conditions:

$$X(t_0) = [S_0, E_0, I_0, Q_0, R_0, D_0], \tag{5}$$

which yields the state space model described as:

$$X_{t+1} = f(X_t) + \xi_t, \tag{6}$$

where f is the nonlinear function, ξ_t is the process noise that is assumed to be Gaussian with zero mean and covariance matrix Ξ. Now, the discrete time nonlinear function $f(X_t)$ can be represented as:

$$f(X_t) = \begin{bmatrix} -\beta IS + \alpha R \\ \beta IS - \epsilon E \\ \epsilon E - \gamma I - qI - dI \\ qI - q_t Q - dQ \\ \gamma I + q_t Q - \alpha R \\ d(N - (S + E + R + D)) \end{bmatrix}. \tag{7}$$

Using the Taylor series approximation to linearize the nonlinear discrete time system in (7) as a linear system we get:

$$F_t = \begin{bmatrix} -\beta I & 0 & -\beta S & 0 & \alpha & 0 \\ \beta I & -\epsilon & -\beta S & 0 & 0 & 0 \\ 0 & \epsilon & (-\gamma - q - d) & 0 & 0 & 0 \\ 0 & 0 & q & (-q_t - d) & 0 & 0 \\ 0 & 0 & \gamma & q_t & -\alpha & 0 \\ -d & -d & 0 & 0 & -d & -d \end{bmatrix}, \tag{8}$$

where F_t is the Jacobian matrix. In the available dataset of reported cases in Saudi Arabia, we use the measurements of the active cases (I) and cumulative death (D) by incorporating them within the model using the measurement equation as:

$$y_t = HX_t + \omega_t, \tag{9}$$

where, y_t is the measurement vector of the observed data and H is the observation matrix structured as:

$$H = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{10}$$

and ω_t is the measurement noise assumed to be Gaussian distributed with zero mean and covariance matrix Ω. The EKF algorithm for state estimation involving both measured and unmeasured states can be described as the following steps:

1) Start with initializing the state vector X_0 and the covariance matrix P_0 as:

$$\begin{aligned} \hat{X_0}^+ &= E[X_0], \\ \hat{P_0}^+ &= E[(X_0 - \hat{X_0}^+)(X_0 - \hat{X_0}^+)^T]. \end{aligned} \tag{11}$$

2) Perform the prediction of state estimates and error covariance as:

$$\begin{aligned} \hat{X}_t^- &= f(\hat{X}_t), \\ P_t^- &= FP_t^+F^T + \Xi_t. \end{aligned} \tag{12}$$

3) Perform the measurement update of the state estimate and estimation error covariance as:

$$\hat{X}_t^+ = \hat{X}_t^- + K_t\left(y_t - H_t\left(\hat{X}_t^-\right)\right), \tag{13}$$

$$K_t = P_t^- H_t^T\left(H_t P_t^- H_t^T + \Omega_t\right)^{-1}, \tag{14}$$

$$P_t^+ = \left(I - K_t H_t\right)P_t^-, \tag{15}$$

where K_t is the Kalman gain and $\hat{+}$ denotes the estimate after processing the measurement whereas $^-$ denotes the process before the correction step.

4) The steps in 2 and 3 can be repeated until getting a better estimate of X_t.

V. Parameter Estimation and Simulation of COVID-19 Spread in Saudi Arabia

We analysed the Saudi Arabia COVID-19 data with active cases (I) and death cases (D) based on the openly available dataset in [20]. The dataset contains daily measurements of these two state variables between the dates 15 February 2020 and 17 March 2022. This dataset contains data about the numbers of tests, cases, deaths, critical cases, active cases and recovered cases in each country. We next used the nested sampling algorithm with Markov Chain Monte Carlo (MCMC) random walk for the live points (N_{live}) to draw samples from the likelihood surface as presented in [19]. The nested sampling is a generic Bayesian inference framework to estimate unknown model parameters along with uncertainty bounds from their posterior probability distribution while also calculating the marginal likelihood or Bayesian evidence (log$Z$) of the model showing the degree of agreement between the model and the measured data. The uncertainty information on the SEIQRD model parameters within the one standard deviation around the mean ($\mu \pm \sigma$) confidence interval (CI) has been shown in Fig. 2. Using the Saudi Arabia COVID-19 data, the estimated model parameters as the mean of the posterior are presented in Table I. The tuning parameters in the nested sampling algorithm to fit the SEIQRD model include: (a) the number of live points $N_{live} = 90$ for exploring the 9D joint posterior distribution of the unknown model parameters, (b) stopping criterion for log-evidence calculation Δlog$Z = 0.01$. As an output we get: (a) total number of likelihood evaluations $N_{like} = 9792$ proportional to N_{live}, (b) model evidence log Z = -7236.54. In the Bayesian inference engine, we used an uninformative prior as a uniform distribution over a specified range of unknown model parameters i.e. $\{\beta, \epsilon, \gamma, d, q, q_t, S_0, E_0, \alpha\}$ as:

$$\begin{gathered} \pi(\beta) \sim \mathcal{U}\left[0, 10^{-4}\right], \quad \pi(\epsilon) \sim \mathcal{U}[0,1], \\ \pi(\gamma) \sim \mathcal{U}[0,1], \quad \pi(d) \sim \mathcal{U}\left[0,1\right], \\ \pi(q) \sim \mathcal{U}\left[0,1\right], \quad \pi(q_t) \sim \mathcal{U}\left[0,1\right], \\ \pi(S_0) \sim \mathcal{U}\left[1 \times 10^5, 1.5 \times 10^9\right], \\ \pi(E_0) \sim \mathcal{U}\left[1 \times 10^5, 1.5 \times 10^9\right], \\ \pi(\alpha) \sim \mathcal{U}\left[0,1\right]. \end{gathered} \tag{16}$$

Using the above prior and a multivariate Gaussian likelihood function assuming the temporal data-points are independent and identically distributed (i.i.d) samples, the nested sampling algorithm draws random samples from the posterior of the unknown SEIQRD model parameters. Fig. 3 shows the posterior distribution of all the parameters of the proposed SEIQRD model with reinfection term using the univariate marginal histograms in the principal diagonal and the bivariate kernel density estimates (KDE) along with scatterplots of the posterior samples in the off-diagonals. Simulation results was conducted using the mean of the posterior prediction to predict the COVID-19 spread in Saudi Arabia as shown in Fig 4. It can be seen from Fig. 4(a) around 150th day of the simulation the infected reported cases reached it peak with 63000 infections

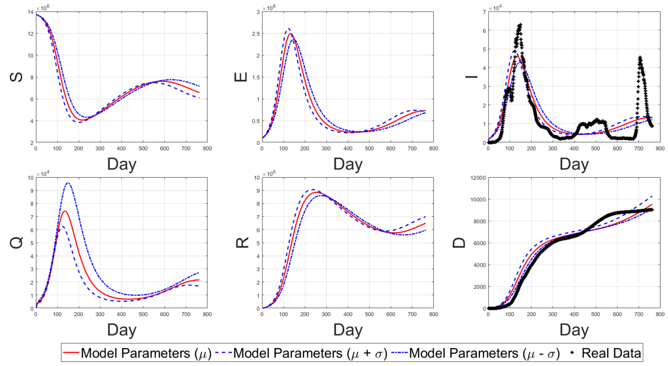

Fig. 2: The estimated uncertainty bounds of the posterior distribution of the model parameters.

and it is noted that the total death in 4(b) in the same period is gradually increasing over time. After the first peak,the number of active cases decreased rapidly which clearly indicates that the Saudi Government has effectively controlled the pandemic using various measures such as lockdown, self-isolation and social distancing which were strictly applied.

For the full period in this study the basic reproduction number R_0 is calculated according to the observed data and estimated around 5.44 for the whole size of the epidemic since the outbreak. In the EKF simulations, several values of the covariances Ξ and Ω were tested and each values has a different performance where we observe that if the Ξ and Ω are small we get a large error in the EKF estimates. This helped deciding to use the process and measurement noise covariance matrices as: $\Xi = 500 \times I_{6\times6}$, $\Omega = diag([100, 1000])$ where we obtained small error in both the infected and death situations which is the best estimate in these cases. Thus, as shown in Fig. 4(a), the measured data of active cases lie very close to the EKF predictions where the mean posterior numerical simulation as the smooth output of the SEIQRD model alone does not perform so well in explaining the non-smooth changes in the active cases. Fig. 4(b) shows the estimated cumulative number of death cases which is around 9000 in Saudi Arabia since the early intervention helped to reduce the mortality rate. The EKF predictions for I and D are very close to the reported data and better than the posterior mean simulations of the SEIQRD model. Each subplot in Fig. 5 corresponds to the simulation of the proposed SEIQRD model and the predictions based on the EKF for the unmeasurable states. Due to the unavailability of the ground truth data for these 4 unmeasurable states, the EKF prediction is more reliable than the smooth dynamical model simulations. However, we notice that the dynamical system simulation model is close to the EKF prediction results in the susceptible cases and recovered cases while for the remaining variables viz. exposed and quarantined cases, the model differs slightly in terms of when they reach their peaks. Thus, the use of the EKF helps in estimating the unmeasurable states such as susceptible, exposed, quarantined and recovered more accurately than the nonlinear system model simulations since it can predict sharp changes while the ODE simulation alone

TABLE I: Mean Posterior of the proposed SEIQRD Model Parameters for the Saudi Arabia COVID-19 Data

Parameter	Value	Description
N	34×10^6	population number
β	2.92×10^{-7}	infection rate
α	0.0028	reinfection rate
ϵ	0.0353	1/incubation period
q	0.9593	quarantine rate
qt	0.5939	time period of quarantine
d	3.5853×10^{-4}	death rate
γ	0.9586	recovery rate
R_0	5.44	Reproductive number

TABLE II: Root Mean Square Error of the EKF-based on SEIQRD Model

Covariance Matrices	Infected Error	Death Error
Ξ = 1 , Ω = diag[10, 10]	196.442	0.74698
Ξ = 0.01 , Ω = diag[10 10000]	212.3171	42.2808
Ξ = 500 , Ω = diag[100, 1000]	56.3065	0.27098

is mostly smooth in nature. In order to quantify the EKF prediction results, we use the root mean square error (RMSE) of the active and deaths cases defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}\left(X_{\text{reported } I,D} - X_{\text{EKF},I,D}\right)^2}, \qquad (17)$$

where, n is the number of measurement points. The RMSE values were compared in Table II, for the two state variables I and D using different assumptions of the noise covariance matrices. This demonstrates the validity and efficacy of the proposed nonlinear state estimation method beside visual comparisons of the real COVID-19 data with the EKF-based predictions.

VI. Conclusion

This paper presents a new epidemiological model of the SEIQRD form with reinfection to understand the impact of COVID-19 based on active and death cases data in Saudi Arabia. Nested sampling algorithm based posterior mean parameters were used in the SEIQRD model for dynamic simulations. The fitted dynamic model can be useful to predict the spread of infectious disease and can be further used to help the Saudi Government to monitor the COVID-19 pandemic since different scenarios of unknown bias/noise covariance have been considered. The EKF was applied with the linearized version of the SEIQRD model to estimate the dynamics of COVID-19 unmeasurable states while also validating the predictions with the actual measurements of the active cases and the cumulative deaths. Our results show that the EKF is capable of estimating the evolution of the pandemic in the long term which yields more accurate estimation than the fitted nonlinear dynamical system model. In the future, we shall consider more complex epidemiological models and other families of the nonlinear Kalman filters with different assumptions of the noise distribution beside the normal case.

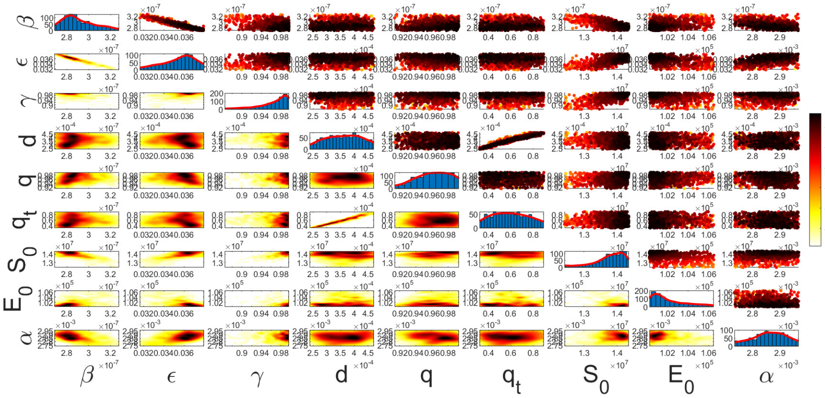

Fig. 3: Posterior distribution of the proposed SEIQRD model parameters with the reinfection term.

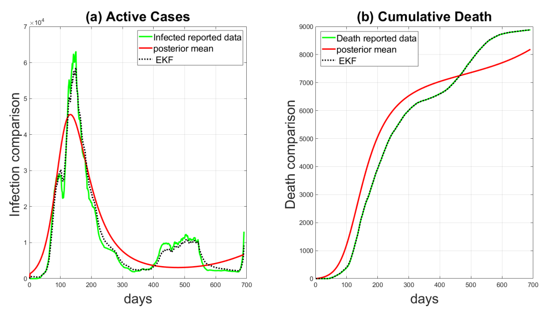

Fig. 4: Comparison of the state estimation based on EKF with real data in Saudi Arabia: (a) active cases (b) cumulative death.

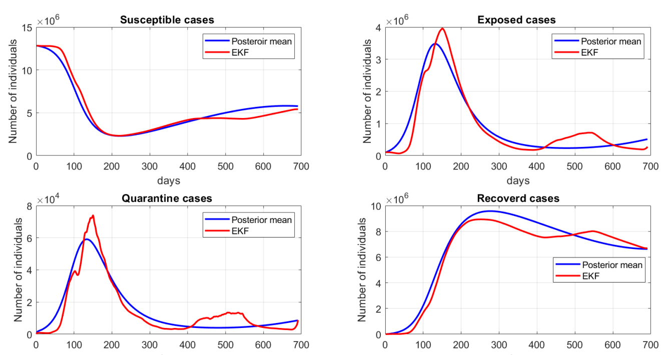

Fig. 5: Comparing estimated variables of the Susceptible, Exposed, Quarantined and Recovered cases in Saudi Arabia.

References

[1] C. Sohrabi, Z. Alsafi, N. O'neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, and R. Agha, "World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19)," *International Journal of Surgery*, vol. 76, pp. 71–76, 2020.

[2] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A*, vol. 115, no. 772, pp. 700–721, 1927.

[3] J. Guan, Y. Zhao, Y. Wei, S. Shen, D. You, R. Zhang, T. Lange, and F. Chen, "Transmission dynamics model and the coronavirus disease 2019 epidemic: applications and challenges," *Medical Review*, vol. 2, no. 1, pp. 89–109, 2022.

[4] P. Birrell, J. Blake, E. Van Leeuwen, N. Gent, and D. De Angelis, "Real-time nowcasting and forecasting of covid-19 dynamics in england: the first wave," *Philosophical Transactions of the Royal Society B*, vol. 376, no. 1829, p. 20200279, 2021.

[5] H. M. Youssef, N. A. Alghamdi, M. A. Ezzat, A. A. El-Bary, and A. M. Shawky, "A modified seir model applied to the data of covid-19 spread in saudi arabia," *AIP advances*, vol. 10, no. 12, p. 125210, 2020.

[6] H. M. Youssef, N. Alghamdi, M. A. Ezzat, A. A. El-Bary, and A. M. Shawky, "A proposed modified seiqr epidemic model to analyze the covid-19 spreading in saudi arabia," *Alexandria Engineering Journal*, vol. 61, no. 3, pp. 2456–2470, 2022.

[7] M. Alauddin, M. A. I. Khan, F. Khan, S. Imtiaz, S. Ahmed, and P. Amyotte, "How can process safety and a risk management approach guide pandemic risk management?" *Journal of Loss Prevention in the Process Industries*, vol. 68, p. 104310, 2020.

[8] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[9] A. Jazwinski, "Stochastic processes and filtering theory. vol. 64 of math," *Science and Engineering. Academic Press. New York*, 1970.

[10] D. Simon, *Optimal state estimation: Kalman, H_∞, and nonlinear approaches*. John Wiley & Sons, 2006.

[11] B. Cazelles and N. Chau, "Using the kalman filter and dynamic models to assess the changing hiv/aids epidemic," *Mathematical Biosciences*, vol. 140, no. 2, pp. 131–154, 1997.

[12] D. Ndanguza, I. S. Mbalawata, H. Haario, and J. M. Tchuenche, "Analysis of bias in an ebola epidemic model by extended kalman filter approach," *Mathematics and Computers in Simulation*, vol. 142, pp. 113–129, 2017.

[13] J. Song, H. Xie, B. Gao, Y. Zhong, C. Gu, and K.-S. Choi, "Maximum likelihood-based extended kalman filter for covid-19 prediction," *Chaos, Solitons & Fractals*, vol. 146, p. 110922, 2021.

[14] X. Zhu, B. Gao, Y. Zhong, C. Gu, and K.-S. Choi, "Extended kalman filter based on stochastic epidemiological model for covid-19 modelling," *Computers in Biology and Medicine*, vol. 137, p. 104810, 2021.

[15] K. K. Singh, S. Kumar, P. Dixit, and M. K. Bajpai, "Kalman filter based short term prediction model for covid-19 spread," *Applied Intelligence*, vol. 51, no. 5, pp. 2714–2726, 2021.

[16] A. S. Breathnach, P. A. Riley, M. P. Cotter, A. C. Houston, M. S. Habibi, and T. D. Planche, "Prior covid-19 significantly reduces the risk of subsequent infection, but reinfections are seen after eight months," *Journal of Infection*, vol. 82, no. 4, pp. e11–e12, 2021.

[17] K. Chatterjee, K. Chatterjee, A. Kumar, and S. Shankar, "Healthcare impact of covid-19 epidemic in india: A stochastic mathematical model," *Medical Journal Armed Forces India*, vol. 76, no. 2, pp. 147–155, 2020.

[18] O. Diekmann, J. Heesterbeek, and M. G. Roberts, "The construction of next-generation matrices for compartmental epidemic models," *Journal of the Royal Society Interface*, vol. 7, no. 47, pp. 873–885, 2010.

[19] J. Skilling, "Nested sampling for general bayesian computation," *Bayesian Analysis*, vol. 1, no. 4, pp. 833–859, 2006.

[20] Worldometer. (2022) Covid-19 worldometer daily snapshots. [Online]. Available: https://www.kaggle.com/datasets/selfishgene/covid19-worldometer-snapshots-since-april-18