GPR and ANN Based Prediction Models for COVID-19 Death Cases

Anwar Jarndal, Saddam Husain, Omar Zaatar, Talal Al Gumaei and Amar Hamadeh
Electrical Engineering Department, University of Sharjah, Sharjah, UAE, ajarndal@sharjah.ac.ae

Abstract — COVID-19 pandemic now affects the entire world and has a major effect on the global economy. A number of medical researchers are currently working in various fields to tackle this pandemic and its circumstances. This paper aims of developing a model that can estimate the number of deaths in the affected cases based on the documented number of older (above 65 years of age), diabetic and smoking cases. The Gaussian Process Regression (GPR) approach has been used to build the model and its performance was compared with a corresponding Artificial Neural Network (ANN) model. The model was applied to reliable data published by the World Health Organization (WHO) for different countries in North America, Europe and the Gulf region. The model provided impressive results with an excellent prediction of data from all the countries under investigation. The model may be useful in estimating the number of deaths due to any arbitrary number of inputs. It would also help to prepare effective measures to minimize the number of deaths.

Keywords—COVID-19, Prediction Model, Gaussian Process Regression.

I. INTRODUCTION

COVID-19 is the infectious disease begot by the most lately discovered coronavirus. It was first identified in Wuhan, China, in December 2019. Due to the surging cases across the globe COVID-19 is classified as pandemic by WHO. As of, 6 June 2020, a total of 11,565,103 cases are reported out of which 6,539,080 are successfully treated. However, unfortunately, 536910 people are succumbed to death by this highly contagious disease [1]. Although, it is regarded in the same family of novel Coronavirus, which also contains infections ranging from common cold to more severe disease such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS), however, the extraordinary ability of COVID-19 to spread through various means within the human beings makes it highly problematic [2]. Moreover, it can be particularly detrimental to the older people, people with weak immunity and those who are suffering from high blood pressure, heart, lung problems, diabetes, and cancer [3]. There are plethora of guidelines issued by various trusted health organizations to contain the virus such as "social distancing", wearing mask, and immediate reporting to the nearest officially recognized healthcare if symptoms were to found until a proper vaccination or treatment is available globally. For those, who are suffering from mild symptoms, self-isolation, self-quarantine, and distancing is recommended. But for those who are suffering from severe symptoms a proper medical care is recommended [4].

There is a growing need for adequate studies and innovative solutions to counter the on-going pandemic. Due to the complex and non-linear nature of the problem, Machine learning (ML) potentially can provide advance and operative models to understand the correlation between various issues pertaining to the disease and its spread. The world communities, such as Kaggle, Our World in data, and WHO, have been very active in gathering the intrinsic data related to COVID-19. These data include case to a case study of various nationalities, number of infected and recovered subjects. Provided with enough data, the ML models can help forecast the likelihood of possible behavior of the COVID-19 pandemic, the number of possible cases in future, and efficient measures to be taken to halt the overall pandemic.

Many papers have been published using ML techniques to develop a forecasting model for COVID-19. One of these papers proposed a model to predict the growth and trend of COVID-19 pandemic based on an ML model and cloud computing [5]. In this work, time-based data has been used for a wide range of countries over the world. Another model based on susceptible-exposed-infected-removed (SEIR) and recurrent neural network (RNN) have been presented in [6]. Both models have been trained on previous data of SARS 2003 and used for prediction in the first months of this year 2020 in China. Also, these models are based on time-based data. ANN model was presented in [7] and also applied on time-based data to predict and forecast the trend of the COVID-19 cases. In [8], a ML forecasting model for the COVID-19 pandemic has been developed and applied on recorded data from India. Different MLs, including ANN, linear regression, and vector auto-regression, have been used to forecast the spread of COVID-19 in India.

Most of the previous models consider time-based data, and these data are mostly monotonic and have deterministic behavior that can be modeled by a quietly simple model. In this paper, a different problem related to COVID-19 will be addressed. This work aims to develop a model that predicts the number of deaths due to COVID-19 based on input information about the number of old cases, diabetic cases, and smoking cases. In principle, these data are not time-based and generally have random behavior. In order to produce the best forecasting model, GPR and ANN supervised based methods are exploited and compared in terms of their prediction abilities. ANN is known as parametric methods, whereas GPR is a probability-based non-parametric method. In this paper, the GPR as an advanced and recently developed ML technique will be demonstrated in predicting the death cases due to COVID-19. GPR performance will be benchmarked against ANN. To the best knowledge of the authors, this paper demonstrates, for the first time, the applicability of GPR for solving such problems.

II. DATASET

The modeling approaches have been applied on daily updated data provided by the WHO situation reports [9]. The dataset is comprehensive and up to date. Three different types of targeted people were used as an input to the model, number of cases above 65 years old, smokers, and diabetic. In addition, patient's data were used to develop three different models in three different regions European region, North America, and the Middle East. Three countries from North America are considered USA, Canada, and Mexico. From

Europe, France, Germany, Italy, Spain, and the United Kingdom are selected. The Middle East list includes Kuwait, United Arab Emirates, Qatar, Saudi Arabia, and Oman.

III. ANN AND GPR MODELING

The figure below shows the proposed architecture of the two models with number of cases above 65 years old, smokers, and diabetic as the inputs and number of deaths as the output.

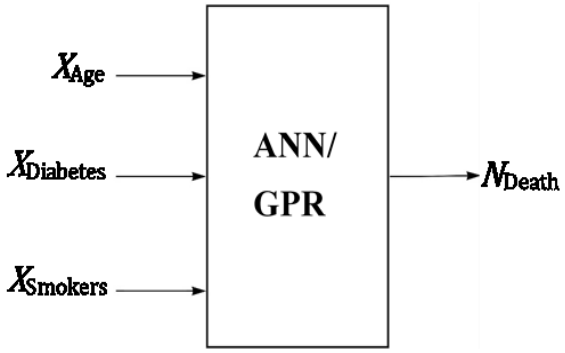

Fig.1: A general structure of ANN/GPR models

A. ANN

The well-formulated ANN model is built to address the correlation between the number of deaths with respect to age, the number of smokers in the given area, and the number of diabetic patients. The ANN model utilized MLP architecture [10]. The model topology is tuned by running the algorithm for different architectures and comparing the MSE and RMSE for the given architecture. To produce the best possible model every single parameter such as pre-processing techniques, activation functions, initialization methods of weights, biases, number of epochs, training, and testing data size, number of hidden layers, number of nodes in a given hidden layer, and training functions are tuned. Analytically, the output, N_{death}, due to COVID is governed by the equation (1).

$$N_{death} = w_b^3 + \sum_{k=1}^{6} w_k \ tanh(w_{kb}^2 + \sum_{j=1}^{6} w_{kj} \ tanh(w_{1j}X_{age} + w_{2j}X_{smokers} + w_{3j}X_{diabetes} + w_{bj}^1)) \quad (1)$$

Where N_{death} is the total number of deaths reported, X_{age} is the number of subject who are aged over 65, $X_{smokers}$ is the number of active smokers and $X_{diabetes}$ is the total number of active diabetic patients. The terms w_{1j}, w_{2j} and w_{3j} are input weights, w_{kj} are the intermediate weights (between two hidden layers). Furthermore, w_{bj}^1, w_{kb}^2 and w_b^3 are the input-layer, hidden-layer and output-layer biases, respectively. The model uses the non-linear $tanh(.)$ activation or threshold function at each layer.

Furthermore, three separate models are built to address, three different regions: European region, Middle East region and North America region. The models topology used to build the ANN is depicted in Table I.

Table I: ANN parameters used to develop the models

Regions	Number of hidden Layers	Number of neurons in each layer	Transfer function at each layer except output	Training function
Eruope	2	6 each	Tan-sigmoid	Bayesian regularization backpropagation
America	2	6 each	Tan-sigmoid	Levenberg-Marquardt backpropagation
Middle East	2	6 each	Tan-sigmoid	Bayesian regularization backpropagation

B. GPR

GPR is a non-parametric kernel-based probabilistic model. Unlike the ANN model, GPR considers the new input vector of the test set and training set to predict the output values for the new test set. Here, instead of finding the weights and biases to fit the best admissible functions, the model directly utilizes the test set. GPR utilizes the Bayesian principle to calculate the posterior distribution by first calculating the prior distribution associated with the data. Due to the page limitation, we are not presenting the mathematical framework in this paper. The entire theoretical and mathematical framework of GPR can be found in [11].

GPR can be easily characterized in terms of mean and covariance. Analytically, Suppose the training set, $\mathcal{D}_t = \{X_t, y_t\}$; where X_t is the multivariate set of inputs composed of training observations of X_{age}, $X_{smokers}$, and $X_{diabetes}$. y_t is the N_{death} . By using the standard results, the conditional distribution $p(f_*|\mathcal{D}_t, T)$ is computed using equation (2) and (3).

$$m(T) = K(T, X_t)[K(X_t, X_t) + \sigma_m^2 I]^{-1} y_t \quad (2)$$

$$k_t(T, T') = k(T, T') - k(T, X_t)[K(X_t, X_t) + \sigma_m^2 I]^{-1} k(X_t, T') \quad (3)$$

Where T is a multivariate set of inputs given in (4). $K(T, T)$ known as Gram matrix, $m(T)$ is the mean, $k_t(T, T')$ is the covariance or kernel function and I denotes the unit matrix.

$$T = \{[X_{age}^i, X_{smokers}^i, X_{diabetes}^i]\}; i = 1,2, \ldots\ldots, n \quad (4)$$

Here, X_{age}^i, $X_{smokers}^i$ and $X_{diabetes}^i$ are the column vectors where each row defines a new observation and n is the total number of observations in the test set.

To predict f_*, it uses simply mean function given in (3) or sample function from the GP with this mean function and kernel (4) as described before. Now, the predicted N_{death} can be analytically calculated using (5).

$$N_{death} = N(m(T), k_t(T, T')) . \quad (5)$$

Like before, three separate models are built to address the different regions. The models were developed using GPR. The dataset is used to train the three models, one for each region. The models were built and optimized using an optimization algorithm. The hyper-parameters are optimized using the Bayesian optimizer by running the algorithm for 100 iterations. All parameters used and the optimal values of the hyper-parameter are presented in Table II.

Table II: GPR parameters used to develop the models

Regions	Sigma	Basis Function	Kernel Function	Fit Method	Active Set Method
Europe	139.98	linear	Rational quadratic	Exact	Random
America	673.86	Constant	Squared Exponential	Exact	Random
Middle East	3.134	Pure Quadratic	rational quadratic covariance	Exact	Random

IV. Simulation Results and Discussion

Both ANN and GPR models were developed to predict the number of deaths due to COVID 19 utilizing the data mentioned in Section II.

A. ANN

The model is developed in MATLAB software environment. First, the dataset is prepared. The data is divided: 84% of the total samples are used for training, and the other 16% are used for validation and testing. It is observed that there is a strong non-linearity in the dataset. Using normalization pre-processing technique, the data is forced to be within the range of -1 to 1, and later, post-processing is done to acquire the previous range. As discussed before, the models are developed using the topology listed in Fig. 1. Due to strong randomness present in the data, ANN could not provide satisfactory results for some cases. The plots of measured and simulated number of deaths for Europe, America, and Middle East regions are depicted in Figs. 2, 3, and 4, respectively. The ANN models are not able to simulate the exact distribution of the data. It can be deducted from the figures that despite ANN's non-linear nature, it could not provide satisfactory results for the European and Middle Eastern regions because of its stronger random behavior. However, it correctly simulated the American region due to the almost linear behavior of the output.

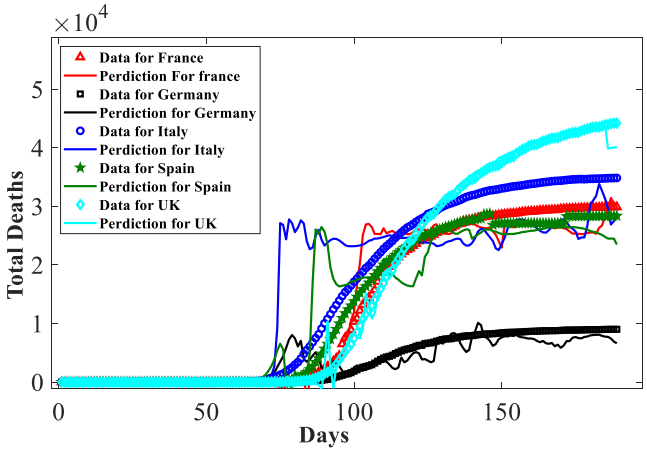

Fig.2. ANN Model Prediction for Europe.

B. GPR

The second model is developed using GPR. The same data is used to build the model. Again, the data is pre-processed before training and later post-processed to return the data into the original format. Likewise, ANN, the data is divided into a 70-30 ratio. 70% of the entire dataset is used for training, and the other 30% to validate and test the model's efficiency. Individual models are built using the parameters listed in Table II for each model. Due to the intrinsic flexibility of GPR to assume the Gaussian probability distribution and fit the model using this presupposition, the GPR has given an excellent performance. To verify the robustness of the model, the same plots are redrawn using GPR models and shown in Figs. 5, 6, and 7.

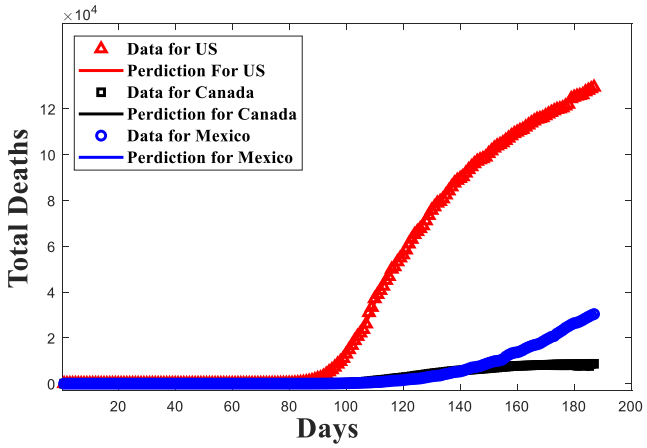

Fig.3. ANN Model Prediction for North America.

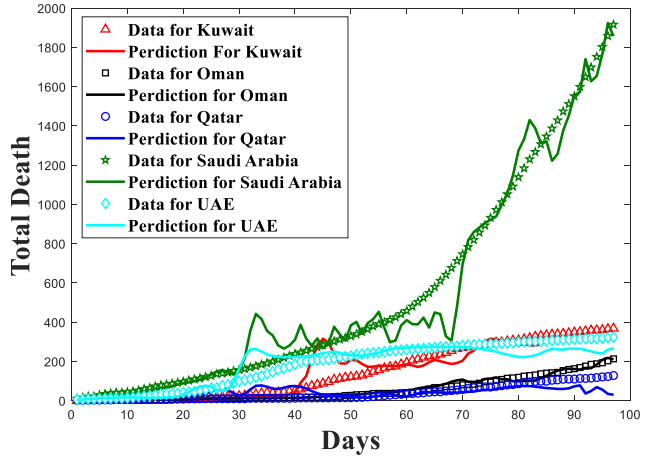

Fig.4. ANN Model Prediction for Middle East

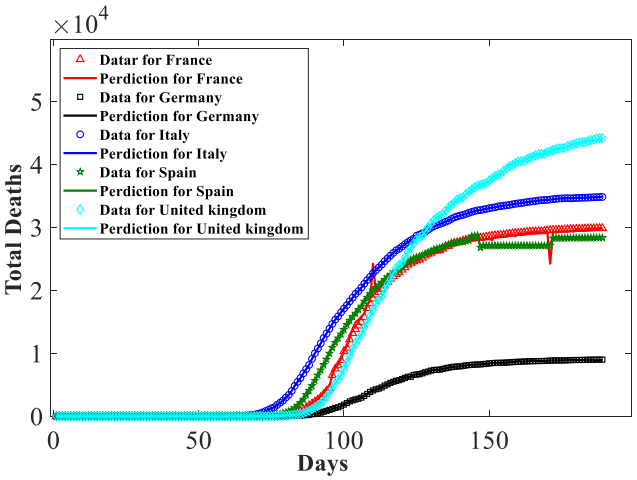

Fig.5. Predict and actual data Europe using GPR model.

It can be observed from the figures the GPR has shown an excellent agreement between the predicted and measured output for all the regions. Furthermore, the MSE, RMSE, and regression results are listed in Table III.

Table III: Performance of the GPR models

Regions	Regression coefficient	MSE	RMSE
Europe	0.99984	4.59E+09	6.78E+04
America	0.999	4.72E+05	686.902
Middle East	1	0.0266	0.1631

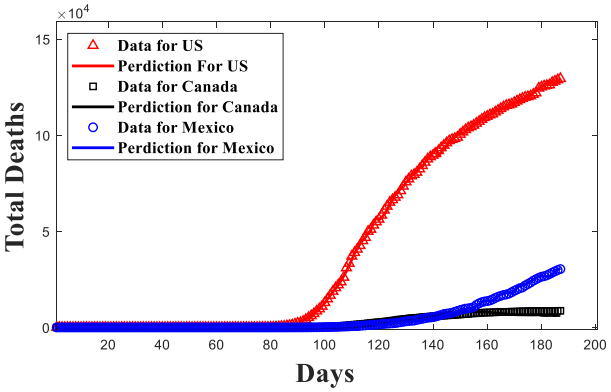

Fig. 6. Predict and actual data for North America using GPR model.

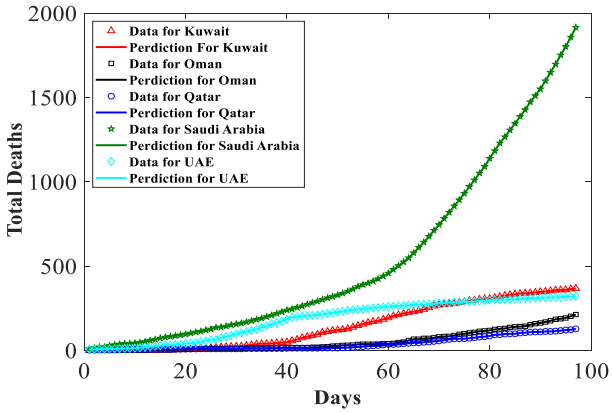

Fig.7. Predict and actual data for the Middle East using GPR model.

To understand the effect of individual factors on the output, individual plots (see Figs. 8-13) are drawn, and analysis is carried out. This is to show the impact level of age, smoking, and diabetic on the number of deaths. These levels could be obtained from the values of the rate of changes as listed in Table IV. As it can be observed, the older people are more impacted by the COVID-19. The results also consistent with the reported data in [12], which show higher rate of death from smoking in USA with respect to.

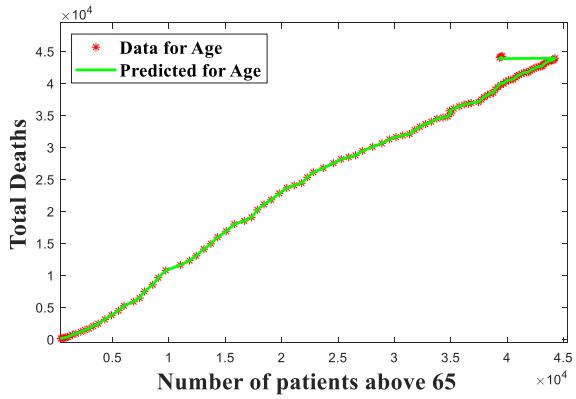

Fig.8. Actual and Predicted Deaths vs Number of patients above 65 (UK) using GPR model

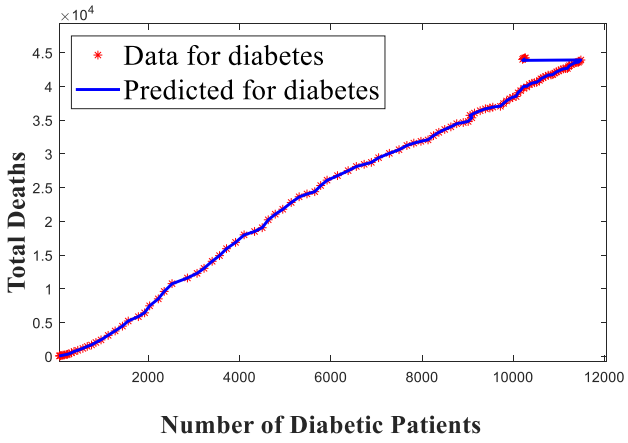

Fig.9. Actual and Predicted Deaths vs Number of diabetic patients (UK) using GPR model.

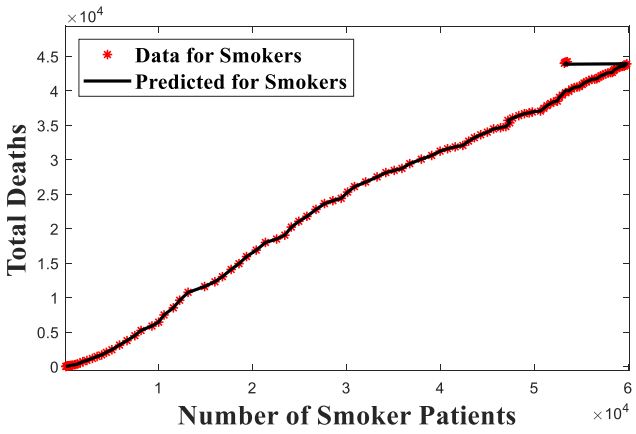

Fig.10. Actual and Predicted Deaths vs Number of smoker patients (UK) using GPR model.

Table IV Rate of change of the output compared to the inputs

Countries	No. of deaths Vs Age	No. of deaths Vs smoking	No. of deaths Vs Diabetic
United Kingdom	1.287	0.509	0.72
USA	0.951	0.639	0.4891

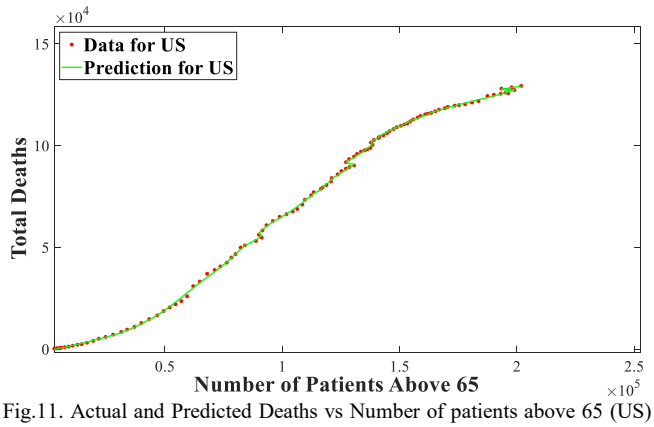

Fig.11. Actual and Predicted Deaths vs Number of patients above 65 (US) using GPR model.

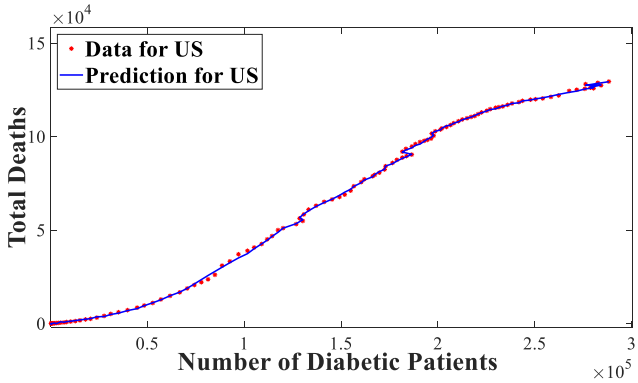

Fig.12. Actual and Predicted Deaths vs Number of Diabetic patients (US) using GPR model.

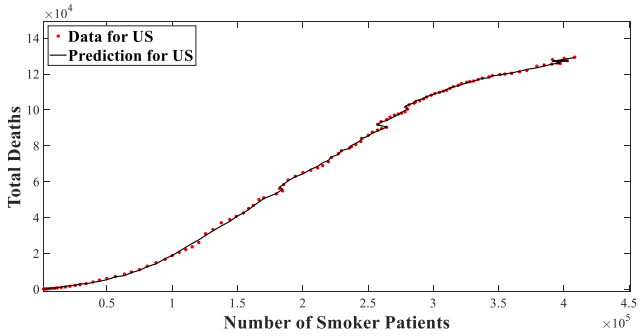

Fig.13. Actual and Predicted Deaths vs Number of smoker patients (US) using GPR model.

The proposed models can contribute to the decision-making of the leaders of these countries. To assess the influence of each factor, it was proposed to measure the rate of change by calculating the increase in the input and the corresponding increase in output. The outcome can be used to evaluate the most influential input relative to the others, thereby giving it more importance. As it was mentioned, our case showed that the Age had the strongest contribution compared to smokers and diabetic patients.

V. CONCLUSION

In this paper, we proposed GPR based model to predict the number of deaths due to the novel COVID-19 and it was benchmarked against ANN model. The impact of age, number of smokers and number of diabetic patients on rising number of deaths due to this highly contagious disease is examined. The data set is taken from WHO, which is publicly accessible to create the models. Apart from using a highly complex model, the ANN-based model could not be adequately capable of predicting the behavior due to the extreme type of randomness present in the data. Provided adequate and balanced data, ANN efficiency can dramatically improve. The efficiency of ANN can be greatly enhanced and could be targeted in a future work. However, GPR has shown substantially improved performance relative to ANN. Thanks to the probabilistic and non-parametric nature of the GPR system, it can be easily simulated and projected. It is found that GPR is more effective than ANN in modeling/ forecasting COVID-19 data.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support from the University of Sharjah, Sharjah, United Arab Emirates.

REFERENCES

[1] Johns Hopkins University CORONAVIRUS RESOURCE CENTER, (https://coronavirus.jhu.edu/).

[2] H. Harpan, et al. ," Coronavirus disease 2019 (COVID-19): A literature review," Journal of Infection and Public Health, Volume 13, Issue 5, May 2020, Pages 667-673.

[3] Gosain, R., Abdou, Y., Singh, A. et al. COVID-19 and Cancer: a Comprehensive Review. Curr Oncol Rep 22, 53 (2020). https://doi.org/10.1007/s11912-020-00934-7.

[4] M. Nicola, " Evidence based management guideline for the OVID19 pandemic - Review article , " International Journal of Surgery, Volume 77, May 2020, Pages 206-216.

[5] S. Tuli , et al.," Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing, " Internet of Things, vol. 11, 2020.

[6] Z. Yang, et al. , "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," J Thorac Dis 2020. doi: 10.21037/ jtd.2020.02.64.

[7] S. Ardabili, " COVID-19 Outbreak Prediction with Machine Learning," medRxiv, April 2020 (doi.org/10.1101/2020.04.17.2007009).

[8] R. Sujath, J. Chatterjee and A.-E. Hassanien, "A machine learning forecasting model for COVID-19 pandemic in India," Stochastic Environmental Research and Risk Assessment, 2020, vol. 34, pp. 959–972.

[9] COVID-19 Open Research Dataset (https://www.kaggle.com/).

[10] Simon Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, 1998.

[11] Rasmussen, C. E. and C. K. I. Williams. Gaussian Processes for Machine Learning. MIT Press. Cambridge, Massachusetts, 2006.

[12] Hannah Ritchie and Max Roser (2019) - "Smoking". Published online at OurWorldInData.org.