2021 Australian & New Zealand Control Conference (ANZCC)
25-26 November 2021. Gold Coast, Australia

Effect of increased number of COVID-19 tests using supervised machine learning models

W. Pooja, N. Snehal, K. Sonam, S. R. Wagh, and N. M. Singh

2021 Australian & New Zealand Control Conference (ANZCC) | 978-1-6654-1650-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ANZCC53563.2021.9628387

***Abstract*—Machine learning is widely being used in medical field for disease diagnostics and research.The area of machine learning is mainly classified into 3 parts: supervised, unsupervised and reinforcement learning.Supervised machine learning (ML) algorithms are used in this paper for modeling and showing the impact of increased testing on the number of daily confirmed cases of COVID-19. The algorithms used to carry out this study are decision tree regression and random forest regression. Machine learning for modeling has proven to be significant for forecasting and hence decision making over the future course of actions. In this paper, Gaussian process regression has been used for modeling as well as forecasting the daily confirmed cases in South Korea. The results obtained show that if the number of tests conducted is increased to the population of South Korea, approximately equal to 51, 286, 183, the peak in the daily cases is obtained earlier and hence the overall number of daily cases is less compared to current cases.**

***Index Terms*—COVID 19, decision tree algorithm, Gaussian process regression, machine learning, random forest algorithm**

I. Introduction

Over the last few years, the popularity of machine learning (ML) and artificial intelligence (AI) has increased vastly due to its ability to efficiently solve complex dynamic system modeling problems. ML is an area of computer science study in which computer algorithms are built to mimic the behavior of humans, like identify existing or prior knowledge, acquire competence and improve its performance with the availability of new data or information. The main categories of machine learning algorithm are supervised learning which builds a model using labeled data where each pair of training data has an input value and its corresponding output, unsupervised learning which trains model using unlabelled data with no output category or label for input values, and reinforcement learning which uses the observations gained from interaction with the environment and the actions taken to reduce the penalty or increase the reward is used to train a model [1].

ML is extensively used in the field of data analysis. The analytical models identify the hidden patterns and features in the data. These patterns and features form the soul of predictive model building. Typically, ML algorithms observe available data, finds underlying features and hidden patterns, and use it to improve its algorithm as new data is made available. The predictive models may either use regression, classification, or clustering based on the available data. A few of the application areas of ML involve predictions in the stock market [2], weather [3], network analysis [4], telecommunication [5], smart grid [6], smart building [7] [8] [9] and healthcare sector [10].

The previous research work shows that the area of ML can be used for prediction of heart disease and detect the transition of agitations [11], chronic kidney diseases and the onset of kidney failure [12], detection of the tumor, and its classification as a benign or malignant [13], etc. Similarly, many machine learning models have been built in the last couple of months to detect SARS coronavirus (SARS-CoV)

W. Pooja, N. Snehal and K. Sonam, S. R. Wagh, and N. M. Singh are with the Control and Decision Research Centre (CDRC), EED, VJTI, Mumbai, India. pkwaghmode_m19@ee.vjti.ac.in, smnaikawadi_m19@ee.vjti.ac.in

also known as Coronavirus disease 2019 (COVID 19) [14] and forecast new cases [15] helping to create awareness among people and make medical facilities and government organizations ready for future possibilities.

As of 6 November 2020, around 52 million people have been infected with corona infection causing the death of around 1.28 million people worldwide. Till the mid-November 2020 no vaccine was invented thus, early diagnosis, isolation, and treatment of infection was the only possible way to control the pandemic caused by a coronavirus [16]. If the whole population including both symptomatic and asymptomatic population is tested for COVID-19 infections [17], the early positive results can be helpful to isolate the patient to reduce the chances of spreading the infection to others [18]. Early treatment reduces complications of the disease and hence the number of resulting deaths.

In this paper a comparison is made for the curve of the number of the confirmed cases of infections plotted against time in the present scenario and the curve obtained when the number of tests carried out is considered to be very large, approximately equal to the population. It has been observed that the increased number of tests would have resulted in an early peak in the curve of daily confirmed cases but the overall number of cases would have been decreased compared to the present scenario. In this paper three machine learning algorithms are used to carry out the study which are as follows:

1) Decision tree algorithm
2) Random forest algorithm
3) Gaussian process regression

Out of these three algorithms, Gaussian process regression is more efficient as it takes model uncertainties into consideration [19].

This paper is structured in seven sections: Section I gives the introduction, in Section II the source of data set and terms in the data set are explained. Different algorithms used in this paper are explained in Section III. The performance parameter, adjusted R^2 score considered to evaluate the model accuracy is given in Section IV. Section V explains the procedure followed in building the model. The results and observations are shown in Section VI and conclusions are put forward in Section VII.

II. Data set

This paper is aimed to study the shift in the number of the day on which the number of positive cases is maximum i.e., the shift in the peak of the curve plotted between the number of days and positive cases. The data used in this paper is collected from the official reports by the 'Our World in Data' [20] team which is updated every day. The data set consists of the time series of the number of tests carried out, the number of positive tests detected, and the number of people who died. The data set is considered during first peak, from the 1^{st} of February 2020 to 4^{th} November 2020.

III. Supervised machine learning algorithms

In this paper, the number of positive patients needs to be predicted for the increased number of tests. Thus, a supervised machine learning model is built. In supervised machine learning, training data has a pair of input features and the corresponding output. There are two approaches for supervised machine learning, regression for continuous data, and classification for discrete data having two or more classes. In this paper, a regressor is used for prediction purposes. The regressor algorithm is trained on a training data set to give a prediction for a given input or test data [21].
Three regression algorithms are considered in this paper:

1) Decision tree regression
2) Random forest regression
3) Gaussian process regression

1. Decision Tree Regression :

The decision tree algorithm breaks the given data into small subsets. An associated decision tree is built incrementally at the same time. The decision tree consists of the decision node, which has two or more branches one for each attribute, and the leaf node, which represents a decision on the numerical target. The decision node at the top of the tree is called the root node. [13] The basic algorithm used for building a decision tree is Iterative Dichotomiser 3 (ID3) [22]. The algorithm starts from the top and starts developing nodes using the greedy approach, i.e., at each iteration,

the best feature at the present moment is chosen to create a node. The best feature is selected by using information gain for each feature. The steps are as shown in fig. When the new input is given, it is

ID3 algorithm

1. Calculate entropy and information gain for each feature
2. Split dataset into subsets using the feature for which the information gain is maximum
3. Make a decision tree node using the feature with maximum information gain
4. If all the rows have same class, current node is made a leaf node with the class as its label
5. Repeat till all the features are covered or decision tree has all the leaf nodes

Fig. 1: Decision tree: ID3 algorithm

assigned to one of the leaf nodes and the predicted output is calculated by taking the average of targets of all the data points falling in that leaf node. This method is highly data sensitive thus there is a high probability of over-fitting.

2. Random Forest Regression :

To overcome the weaknesses in the decision tree algorithm, an ensemble learning algorithm called a random forest algorithm is used. It is a bagging technique that involves random sampling of small subsets of data from the data set. It constructs a multitude of decision trees during the training phase [23]. The output obtained is the mean prediction of all the individual trees. To prevent high co-relation between different trees, some modifications are done, which are as follows:

1) Limit the number of features to be split on at each node to some percentage of total features. These features are called hyperparameters.
2) During the split generation phase, each tree draws random samples from the original data set to prevent overfitting.

3. Gaussian Process regression :

Gaussian process regression also called Kriging, is a non-parametric algorithm that calculates the probability distribution over all possible functions fitting a given data using Bayes' formula [21] [9].

$$posterior = \frac{likelihood \;\; X \;\; prior}{marginal \;\; likelihood} \tag{1}$$

In Gaussian process regression (GPR), a Gaussian process prior is assumed which is given by the equation [21]:

$$f(x) \sim GP(m(x), K(X, X')) \tag{2}$$

where,
m(x) is a mean function which is usually considered to be 0 and
K(X, X') is a co-variance function.
The co-variance function or kernel matrix is a positive semi-definite matrix. The most commonly used kernel function is a squared exponential matrix which is given as follows [21]:

$$K(X, X') = \sigma_f^2 \;\; exp(-\frac{1}{2l^2}||X - X'||^2) \tag{3}$$

In the above equation the variance parameter, σ^2, and length scale, l are the parameters that control the distribution of model parameters. These parameters are called hyperparameters [24]. Initial values of these parameters are assumed in prior and are later optimized using different techniques like marginal likelihood optimization. The prior distribution of these hyperparameters does not affect their optimized values and thus the performance of GPR.

The independently identical distributed (i.i.d.) Gaussian noise $\epsilon \sim (o, \sigma^2)$ is incorporated in prior as:

$$y \sim GP(m(x), K(X, X') + \delta_{ij}\sigma_n^2) \tag{4}$$

where, δ_{ij} is a Kronecker delta whose value is 1 for $i = j$ and 0 otherwise.

The data set is split into two subsets, training subset and test subset. The training subset is used for training the model and the test subset is used for validation purposes. The joint multivariate Gaussian distribution of training data and test data is obtained from the Gaussian prior as [21]:

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim N(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}) \tag{5}$$

where, f_* is a posterior for test data.

To predict the posterior distribution, the data and test observations are conditioned out from the posterior distribution. The posterior distribution obtained is a Gaussian distribution with mean and co-variance which are given as [21]:

$$f^*|X, y, X^* \sim N(\overline{\mathrm{f}^*}, \Sigma^*) \tag{6}$$

$$\overline{\mathrm{f}^*} = \mu^* + K(X^*, X)[K(X, X) + \sigma_n^2 I]^{-1}(y - \mu) \tag{7}$$

$$\Sigma^* = K(X^*, X^*) - K(X^*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X^*) \tag{8}$$

IV. Adjusted R^2 score ($R^2_{adjusted}$)

$R^2_{adjusted}$'s value gets adjusted with the addition of new features in the data set. It increases if the features added are useful.Greater the score of $R^2_{adjusted}$, better is the model improvement. It is given by the equation:

$$R^2_{adjusted} = 1 - (1 - R^2)\frac{(n-1)}{n-(k+1)} \tag{9}$$

where n and k denotes the sample size and the number of independent variables respectively in the regression equation.

V. Procedure

The data set used in this paper has a time series information about the number of tests conducted, the number of positive cases and the number of people who died. Most of the data available online may have some data points missing thus, data preprocessing is required to be done before it is fed to an algorithm. The cleaned data is then divided into two parts: training data (around 80 percent) and test data (around 20 percent). The training data is used to train the model which is further validated using test data. In this paper, three algorithms have been used which are decision tree regression, random forest regression and GPR. The resulting curves are included in the observations and results section.

VI. Results and observations

Figure 2 and 3 shows the curve fitting to the number of daily positive cases of COVID-19 in South Korea from duration mentioned above using decision tree regression and random forest regression respectively. Here it can be observed that the peak of positive cases of coronavirus from South Korea is on 60^{th} day. Figure 4 and 5 gives the curve fitting to the number of daily positive cases of COVID-19 in South Korea when the number of daily tests is equal to population using deci-

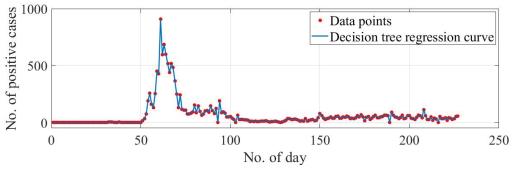

Fig. 2: Plot of Number of day vs number of positive cases for COVID-19 in South Korea using decision tree regression.

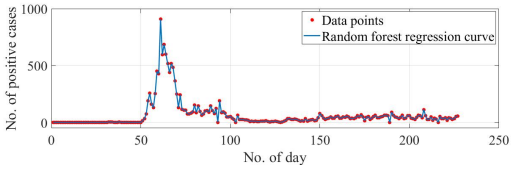

Fig. 3: Plot of Number of day vs number of positive cases for COVID-19 in South Korea using random forest regression.

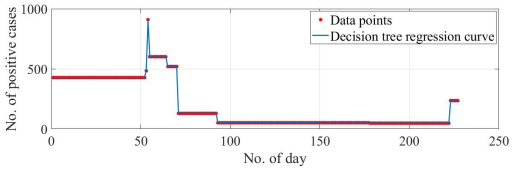

Fig. 4: Plot of Number of day vs number of positive cases for COVID-19 in South Korea when no of tests done daily are equal to the population of South Korea using decision tree regression.

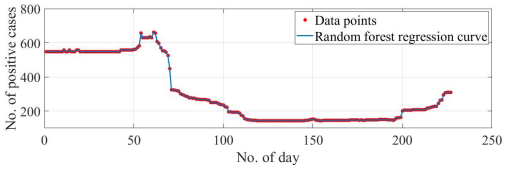

Fig. 5: Plot of Number of day vs number of positive cases for COVID-19 in South Korea when no of tests done daily are equal to the population of South Korea using random forest regression.

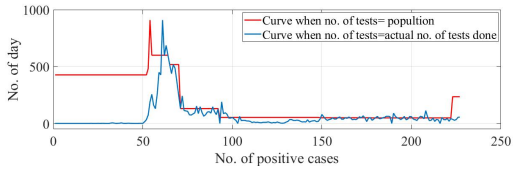

Fig. 6: Peak Comparison plot for number of positive cases when tests conducted are equal to population and existing data of conducted tests using decision tree regression.

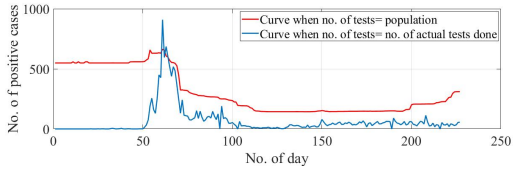

Fig. 7: Peak Comparison plot for number of positive cases when tests conducted are equal to population and existing data of conducted tests using random forest regression.

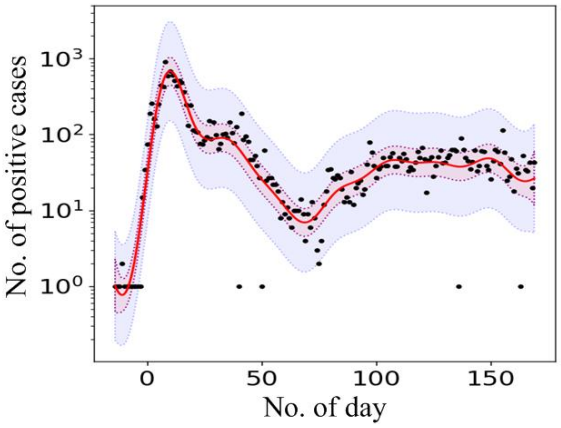

Fig. 8: Plot of Number of day vs number of positive cases for Covid-19 in South Korea using Gaussian process regression.

sion tree regression and random forest regression respectively. In Figure 6 and 7, it can be clearly observed that when number of tests conducted daily in South Korea for COVID-19 are equal to the population of South Korea, the peak for positive cases shifts and occurs on 54^{th} day. The $R^2_{adjusted}$ values for the decision tree regression model and random

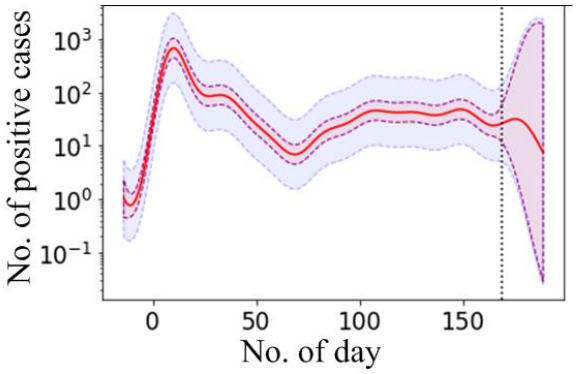

Fig. 9: Plot of prediction of number of positive cases for Covid-19 in South Korea using Gaussian process regression.

forest regression model obtained during experiment are 0.6732853471885556 and 0.8537418153934455 respectively. Since $R^2_{adjusted}$ value of random forest regression model is greater than that of decision tree algorithm, it can be concluded that random forest regression is comparatively accurate model.

In Figure 8, Gaussian process regression result for COVID-19 positive cases in South Korea is shown.Here features considered are number of daily deaths due to COVID-19 and number of daily positive cases for COVID-19 in South Korea. Hyper parameter values for number of positive cases are amplitude is 3.295798, length scale is 12.983654, observational noise variance is 0.189938. The red line in Figure 8 is mean, the data points are indicated by black dots, the blue portion is the variance region with noise and purple portion is the variance region without noise.

The plot of number of days and daily positive cases with forecasting for 20 number of days is shown in Figure 9. The variance band is thick for the days on which the prediction is made. Increased variance shows the flexibility of the model and thus its high efficiency as the uncertainty comes in picture.

VII. Conclusion

The COVID-19 pandemic can be responsible for a massive global crisis. In this study, an ML-based prediction system has been proposed for predicting the number of daily positive cases in South Korea

if the number of tests done for the COVID-19 is equal to the population of South Korea. In that case peak of the positive cases occurs earlier. The study can also be more helpful to authorities to take correct decisions and timely actions to control the COVID-19 crisis. The study will be enhanced in the future, exploring the prediction methodology with an updated data set. The Gaussian process will be more helpful for future work as uncertainties are considered in it.

REFERENCES

[1] C. M. Bishop, "Pattern recognition," *Machine learning*, vol. 128, no. 9, 2006.

[2] K. Pahwa and N. Agarwal, "Stock market analysis using supervised machine learning," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 197–200, 2019.

[3] S. E. Haupt, J. Cowie, S. Linden, T. McCandless, B. Kosovic, and S. Alessandrini, "Machine learning for applied weather prediction," in *2018 IEEE 14th International Conference on e-Science (e-Science)*, pp. 276–277, 2018.

[4] D. Cote, "Using machine learning in communication networks [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 10, pp. D100–D109, 2018.

[5] I. Firdausi, C. lim, A. Erwin, and A. S. Nugroho, "Analysis of machine learning techniques used in behavior-based malware detection," in *2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pp. 201–203, 2010.

[6] K. Sunny, A. Sheikh, and S. Bhil, "Forecasting and enhancing the performance of the electric grid: A dynamic mode decomposition approach," in *2020 7th International Conference on Control, Decision and Information Technologies (CoDIT)*, vol. 1, pp. 1191–1196, IEEE, 2020.

[7] K. Sunny, A. Sheikh, S. Wagh, and N. Singh, "Prediction and classification of temperature data in smart building using dynamic mode decomposition," in *2020 28th Mediterranean Conference on Control and Automation (MED)*, pp. 1074–1079, IEEE, 2020.

[8] K. Sunny, A. Sheikh, and S. Wagh, "Application of dynamic mode decomposition for temperature analysis in smart building," in *2020 7th International Conference on Control, Decision and Information Technologies (CoDIT)*, vol. 1, pp. 1197–1202, IEEE, 2020.

[9] G. Revati, J. Hozefa, S. Shadab, A. Sheikh, S. Wagh, and N. Singh, "Smart building energy management: Load profile prediction using machine learning," in *2021 29th Mediterranean Conference on Control and Automation (MED)*, pp. 380–385, IEEE, 2021.

[10] K. Yazhini and D. Loganathan, "A state of art approaches on deep learning models in healthcare: An application perspective," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 195–200, 2019.

[11] G. E. Sakr, I. H. Elhajj, and H. A. Huijer, "Support vector machines to define and detect agitation transition," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 98–108, 2010.

[12] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *2016 Management and Innovation Technology International Conference (MITicon)*, pp. MIT–80–MIT–83, 2016.

[13] Liyang Wei, Yongyi Yang, R. M. Nishikawa, and Yulei Jiang, "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," *IEEE Transactions on Medical Imaging*, vol. 24, no. 3, pp. 371–380, 2005.

[14] B. Allanach, T. Baldauf, H. Banks, S. Crew, J. Davighi, W. Haddadin, M. Madigan, M. McCullough, C. Turner, and M. Ubiali, "Modelling covid-19 infection data with a simple gaussian process," 2020.

[15] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B. On, W. Aslam, and G. S. Choi, "Covid-19 future forecasting using supervised machine learning models," *IEEE Access*, 2020.

[16] H. Li, S.-M. Liu, X.-H. Yu, S.-L. Tang, and C.-K. Tang, "Coronavirus disease 2019 (covid-19): current status and future perspective," *International journal of antimicrobial agents*, p. 105951, 2020.

[17] P. Gautret, R. Charrel, K. Belhouchat, T. Drali, S. Benkouiten, A. Nougairede, C. Zandotti, Z. Memish, M. Al Masri, C. Gaillard, *et al.*, "Lack of nasal carriage of novel corona virus (hcov-emc) in french hajj pilgrims returning from the hajj 2012, despite a high rate of respiratory symptoms," *Clinical Microbiology and Infection*, vol. 19, no. 7, pp. E315–E317, 2013.

[18] D. Lupi, B. Binda, F. Montali, A. Natili, L. Lancione, D. Chiappori, I. Parzanese, D. Maccarone, and F. Pisani, "Transplant patients' isolation and social distancing because of covid-19: analysis of the resilient capacities of the transplant in the management of the coronavirus emergency," in *Transplantation proceedings*, Elsevier, 2020.

[19] S. Shadab, J. Hozefa, K. Sonam, S. Wagh, and N. M. Singh, "Gaussian process surrogate model for an effective life assessment of transformer considering model and measurement uncertainties," *International Journal of Electrical Power & Energy Systems*, vol. 134, p. 107401, 2022.

[20] E. O.-O. Max Roser, Hannah Ritchie and J. Hasell, "Coronavirus pandemic (covid-19)," *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

[21] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer school on machine learning*, pp. 63–71, Springer, 2003.

[22] H. Zhang and R. Zhou, "The analysis and optimization of decision tree based on id3 algorithm," in *2017 9th International Conference on Modelling, Identification and Control (ICMIC)*, pp. 924–928, 2017.

[23] Y. Li, S. Wang, and X. Ding, "Person-independent head pose estimation based on random forest regression," in *2010 IEEE International Conference on Image Processing*, pp. 1521–1524, 2010.

[24] S. Sundararajan and S. S. Keerthi, "Predictive approaches for choosing hyperparameters in gaussian processes," *Neural Computation*, vol. 13, no. 5, pp. 1103–1118, 2001.