Amir Nasir[1], Seyed Vahab AL-Din Makki[1], Ali Al-Sabbagh[2,3]

Razi University, Kermanshah, Iran (1), Al Taff University College, Kerbala, Iraq (2), Ministry of communication, ITPC , Iraq (3)
ORCID: 1.0009-0002-2596-4685; 2.0000-0001-7125-8357; 3.0000-0003-3270-980x

doi:10.15199/48.2024.05.39

Pandemia Prediction Using Machine Learning

***Abstract**. Coronavirus disease 2019 (COVID-19) is caused a large number of death. Therefore, that Artificial Intelligence (A.I) solution might be capable to identify COVID-19 quickly and early. This paper applies Three ML models to Covid-19 prediction process. We discovered the main dominant variable to decide the negative or positive patient by using different ML models in the prediction process, for instance (LR, XG Boost, and RF). The study and models have been applied for one million patients from European Commission (EC), this data set (cough, fever, sore throat, breath, and headache) been considered as a data sensor coming to the proposed system. The aim is to choose the best ML model for Covid-9 prediction. In addition, all models and dataset have been sufficiently presented with all clarifications and justifications. Also, our data have been provided for one million patients from European Commission (EC). Then, feature selection to prepare the dominant parameters of Cvid-19, which are (cough, fever, sore throat, breath, and headache). As a result, the RF and XG boost obtained the best accuracy in the decision of positive or negative based on nine variables.*

Streszczenmie. *Choroba koronawirusowa 2019 (COVID-19) jest przyczyną dużej liczby zgonów. Dlatego to rozwiązanie oparte na sztucznej inteligencji (AI) może być w stanie szybko i wcześnie zidentyfikować Covid-19. W artykule zastosowano trzy modele ML do procesu przewidywania Covid-19. Odkryliśmy główną dominującą zmienną decydującą o tym, czy pacjent jest negatywny, czy pozytywny, stosując w procesie przewidywania różne modele ML, na przykład (LR, XG Boost i RF). Badanie i modele zastosowano w przypadku miliona pacjentów z Komisji Europejskiej (KE). Ten zestaw danych (kaszel, gorączka, ból gardła, oddech i ból głowy) uznano za czujnik danych docierających do proponowanego systemu. Celem jest wybór najlepszego modelu ML do przewidywania Covid-9. Ponadto wszystkie modele i zbiory danych zostały dostatecznie przedstawione ze wszystkimi wyjaśnieniami i uzasadnieniami. Nasze dane dotyczące miliona pacjentów przekazała także Komisja Europejska (KE). Następnie dokonaj selekcji cech, aby przygotować dominujące parametry Cvid-19, którymi są (kaszel, gorączka, ból gardła, oddech i ból głowy). W rezultacie wzmocnienie RF i XG uzyskało najlepszą dokładność w podejmowaniu decyzji pozytywnej lub negatywnej na podstawie dziewięciu zmiennych.* (***Przewidywanie pandemii za pomocą uczenia maszynowego***)

Keywords: COVID-19, Intelligent framework, Smart detection, Machine Learning
Słowa kluczowe: COVID-19, Inteligentne środowisko, Inteligentne wykrywanie, Uczenie maszynowe

Introduction

A rapid emergence of the novel corona virus (SARS-CoV-2), towards the end of 2019, led to the COVID-19 global epidemic in 2020. Industry, government, and academia from each nation are working together as of May 2020 to discover therapeutic medications and take action to stop the spread of the COVID-19 illness. The corona virus has threatened humans three times in the past century. SARS in 2003, MERS in 2012, and COVID-19 are the three most recent outbreaks. SARS, MERS, and COVID-19 comparison [1] [2][3]. Fever, coughing, sore throat, headaches, exhaustion, muscle aches, and shortness of breath were among the signs of COVID-19. As previously noted, polymerase chain reaction testing is one of the most popular methods for identifying COVID-19 in people. However, diagnosing people using this method takes time, and the outcomes have a high level of false-negative mistakes. [4][5] [6] It may be argued that computer science, particularly machine learning (ML) and artificial intelligence (AI), has advanced significantly. In reality, a variety of computer science techniques have been created to stop the COVID-19 infection from propagating. Infection spreading analysis, drug development support, automatic diagnosis, diagnosis support, social trend analysis, and infection route analysis are a few examples of how AI and ML are employed [1]. Three methods were utilized in this study. The first was the Extreme Gradient Boosting (XGBoost) strategy, which is a scalable machine learning system for tree boosting and It has demonstrated to be an effective and capable machine learning problem solver. A common technique for creating a forecasting model and a quantified boosting algorithm is gradient boosting. It was first created in 2011 by Chen Tianqi and Carlos Gestrin, and several scientists have since refined and improved it [7][8] [9]

Second, Breiman's random forest method uses a machine learning system that has numerous decision trees. It combines the Bagging and Random Subspaces approaches. This technique, one of the greatest machine learning algorithms utilized in many different disciplines, has recently demonstrated its effectiveness in both regression and classification issues. In the RF algorithm, the data set is separated into training and validation (the-out-of-bag) data at random in order to test the learning level. In the data set, training data make up two thirds while validation data make up the other third. The data collection is then used to generate "boot-strap samples" for a large number of decision trees. [10] [11] [12] Third, a logistic regression analysis model explains the relationship on a continuous or categorical scale between the predicted variable, which has two or more categories, and one or more independent variables. Similar methods and processes are used in the logistic regression method and the linear regression method. The logistic regression approach calculates the parameter values using the maximum likelihood estimation (MLE) method, whereas the linear regression method frequently employs the ordinary least squares (OLS) method. [13] [14] [15] Additionally, the dataset and all models have been adequately presented with all reasons and clarifications. Additionally, the European Commission (EC) requested our data for a million patients, and this data set will be regarded as a data sensor for the suggested system. Following the removal of outliers from the data, features will be chosen to prepare the dominating Cvid-19 parameters (cough, fever, sore throat, breath, and headache). Data modelling and visualization are the third phase. Figure 1 illustrates the contribution of this work.

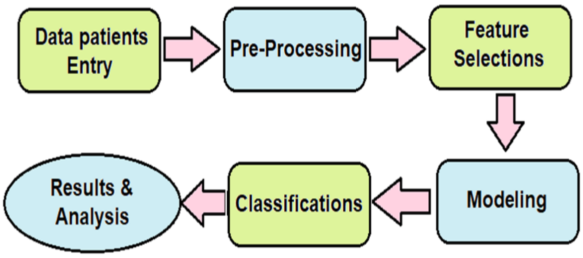

Fig.1 1 million patients, process, modelling, and 3 methods results and analysis.

2. Methodology

An innovative approach are required to create, manage, and analyse big data on the expanding network of infected individuals, patient information, and their community movements, as well as to combine these data with clinical trials, pharmaceutical, genomic, and public health data. In order to analyse the growth of infection with community behaviour, many different sources of data, such as text messages, online communications, social media, and Web articles, might be very beneficial. Researchers can predict where and when the disease is likely to spread using this data combined with machine learning (ML) and artificial intelligence (AI), and they may alert those places to make the necessary preparations. [16] [17] [18].

2.1 XGBoost model:

The Extreme Gradient Boosting (XGBoost) method is a distributed system that has been optimized for tree boosting. It has demonstrated to be an expert and capable problem solver for machine learning Combining a large number of low-accuracy prediction models into a single high-accuracy model is the primary idea behind boosting (improving machine learning models) [7]. For objective function optimization, the XGBoost algorithm employs an adaptive training approach. As a result, the outcome of each stage in the optimization process depends on the prior phase. The following is a list of the XGBoost model's objective function's mathematical expression:

(1) $F_o^i = \sum_{k=1}^{n} l(y_k, y_k^{i-1} + f_i(x_k)) + R\,(f_I) + C$

where the constant term is C, the t-th iteration loss term is provided as l, and the model's regularization parameter R is further defined as:

(2) $(f_i\,) = \gamma\, T_i\, \sum_{j=1}^{T} \omega_j^2$

In general, the value of the g and l customisation options has a direct relationship with how simple the tree structure is. The parameter values have a linear relationship with the simplicity of the tree structure. The model's first and second derivatives, g and h, are provided as follows:

(3) $g_j = \partial_{\hat{y}_k^{i-1}} l\,(\,y_j\,, \hat{y}_k^{i-1})$

(4) $h_j = \partial^2_{\hat{y}_k^{i-1}} l\,(\,y_j\,, \hat{y}_k^{i-1})$

(5) $\omega_j^* = -\frac{\sum g_t}{\sum h_t + \lambda}$

(6) $F_o^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum g)^2}{\sum h + \lambda} + \gamma T$

The formulas listed below are used to find the answer: where the solution weights are given as wj and the loss function score is given as Fo . [19] [20]

2.2 Random forest model

A machine learning system called a Random Forest is made up of several decision trees. It builds each individual tree using bootstrap and feature randomness to produce an uncorrelated forest that has a better forecast than any single tree. The following is a representation of the algorithm for creating a Random Forest made up of N trees: For each time n = 1,.., N: then: create an example using bootstrap, Xn. Create a decision tree using sample Xn and the sample bn. Select the best attribute based on the criteria provided, split the tree based on that split, and repeat this process until the sample is used up. The tree is constructed up to a specified height or until each leaf contains no more than nmin items. Find the best way to separate each partition using m random features chosen from n beginning characteristics. This is how the finished regression algorithm appears:

(7) $f\,(x) = \frac{1}{N} \sum_{I=1}^{N} b_i\,(x)$

where bi(x) is a regression tree. In regression tasks, m = n/3, where n is the total number of features, is the suggested number of random features. The following steps must be taken in order to increase the Random Forest method's predicting accuracy: Possess characteristics with some predictive ability, and Predictions about uncorrelated forest trees. A proper feature and hyperparameter selection while creating weak correlations. The random subspace approach lessens tree correlation and prevents overfitting. On diverse, randomly chosen subsets of the feature description, the fundamental algorithm is taught. By merging the posterior probabilities, it is important to combine the outcomes of many L models [21] [22].

2.3 Logistic Regression model

One way to describe the link between a continuous or category scale's predicted variable, which has two or more categories, and one or more independent variables is to use a logistic regression analytic model. [13] Among generalized linear models (GLM), the logistic regression model is one. Given by is the logistic regression model.

(8) $P_i = f\,(y\,|x\,) = \frac{1}{1+\exp\{-\,(\beta_o + \beta_1 x_i)\}}$

Where Pi is the probability of success and , $P\,(\,y_i = 1) = P_i \; and \; P(y_i = 0) = \; q_i = 1 - P_i\,, 0 \; \le P_i \le 1$. Additionally, the model's parameters are b0 and b1, whereas xi is an independent variable and exp is the mathematical constant known as Euler's number, which roughly equals 2.78. Multiple independent variables, which can be continuous or categorical, can be included in a logistic regression. The following is a form of the multiple logistic regression models:

(9) $P_i = \frac{1}{1+\exp\{-z_i\}} = \frac{\exp\{z_i\}}{1+\exp\{z_i\}}$

Where, $z_i = \beta_0 + \; \beta_1 x_1 + \; \beta_2 x_2 + \cdots + \; \beta_n x_n$, and $\frac{p_i}{1-pi} = \frac{1+\exp\{z_i\}}{1+\exp\{-z_i\}} = \exp\{z_i\}\,.Where.\frac{p_i}{1-p_i}$ is the odd ratio defined as the probability of occurrence the event divided by the prop ability of not occurrence the event. The odd ratio is a solution Of the upper and lower limits for propability where, $0 \; < \text{odd ratio} \; < \infty$. [23]

(10) $p_i = \frac{odds}{1+\,odds} = \frac{exp\,\{\beta_0 + \,\beta_1 x_1 + \,\beta_2 x_2 + \ldots + \,\beta_n x_n\}}{1\,+\,exp\;\{\beta_0 + \,\beta_1 x_1 + \,\beta_2 x_2 + \ldots + \,\beta_n x_n\}}$

(11) logit = In (odds) = In ($\frac{P_i}{1-\,P_i}$) = z_i

(12) logit = $\beta_0 + \; \beta_1 x_1 + \; \beta_2 x_2 + \ldots + \beta_n x_n, - \infty \; < \; lpgit \; < \infty$

3. Dataset

The dataset have been used in this research about one million patients from European Commission (EC) from 20 March 2020 to 11 Oct 2021[24]. The original data as a csv format and includes eight parameters as (cough, fever, sore throat, breath, gender, age, headache, test results). This data also includes female 53% and 12% above 60 years. Figure 2 shows the correlation matrix for each variables related to each other. It presents how each variable will effect on the other variables such as cough or fever and so on in the data set that used in this work. In this section, we presents the results three prediction models of Covid-9 based on dataset. After data preparation, The collected datasets were divided into training and testing sets to facilitate the Covid-19 modeling process. The datasets are randomly shuffled as training and testing sets to eliminate data bias and generate a more accurate model compared to the three models that explained in earlier in this paper. Training set is utilized for optimization of learning model, and the testing set used to assess the accuracy of the used models (70% and 30% respectively). To analyze the performance of ML models efficiently we used a cross validation in testing and validation.

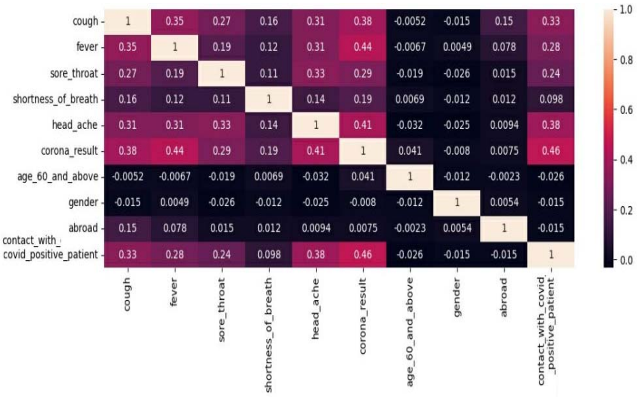

Fig.2 Matrix correlation.

4. Results and discussion

In this section, we present the results three prediction models of Covid-9 based on dataset. After data preparation, The collected datasets were divided into training and testing sets to facilitate the Covid-19 modeling process. The datasets [24] are randomly shuffled as training and testing sets to eliminate data bias and generate a more accurate model compared to the three models that explained in earlier in this paper. Training set is utilized for optimization of learning model, and the testing set used to assess the accuracy of the used models (70% and 30% respectively). To analyze the performance of ML models efficiently we used a cross validation in testing and validation. The following figures will explain in details to clarify the results

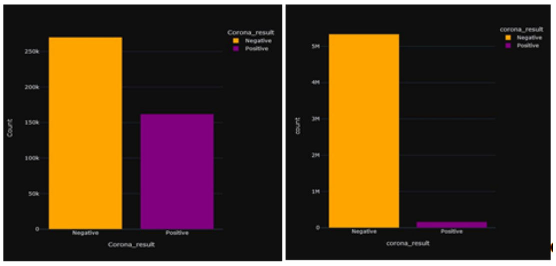

Fig.3 a- without under sampling, b- with under sampling

Figure (3-a) shows clearly the distributed number of patients (negative "orange") and non-patients (positive "purple"). This is very important chart to avoid the rare pattern to be sure that the training was correctly done or not. Based on the big difference between two categories (negative and positive), so the next step is under sampling have to be done to make sure the process in the correct way. Figure (3-b) presents the results after under sampling issue. This clearly shows that the positive is about half of the negative issues, so now the classification models are working efficiently.

In other hand, figure 3-b after feature selection explains the three methods that used in this research. It shows clearly the importance of each selected variable from dataset to decide the patient's test is negative or positive. Figure 4 presents the importance variables in descent for each method as follow: logistic regression (in top), the random forest (in middle), and the XG-boost (in bottom). This figure in all shows that the contact, fever, headache, and cough are the main variables score for decision making (patient or non-patient).

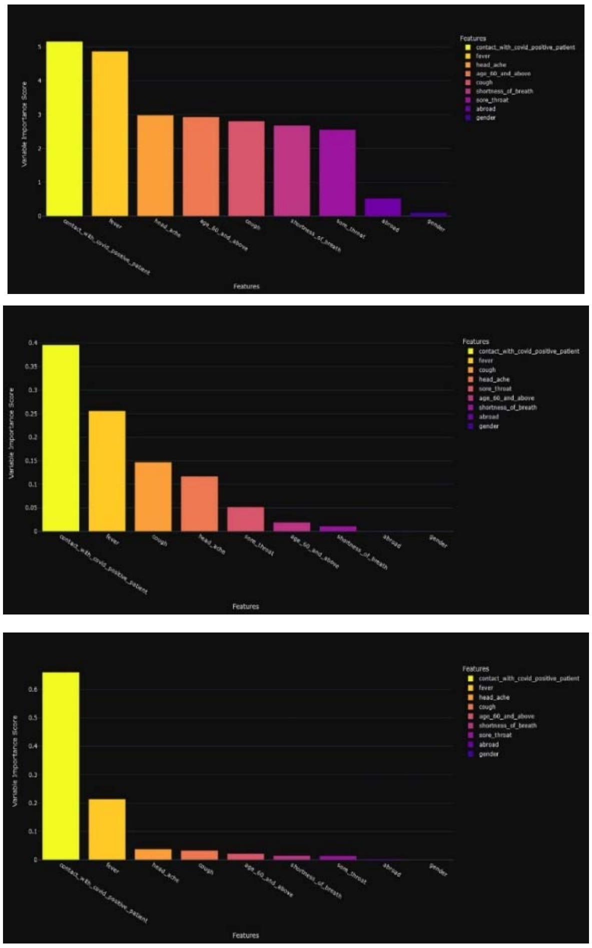

Fig. 4 LR, RF, and XG boost performance.

5. Conclusion

As known, COVID-19 is caused a large number of death since has declared an international pandemic in December 2019 and it is spreading all over the world. This paper applies ML models to Covid-19 prediction process. We discovered the main dominant variable to decide the negative or positive patient by using different ML models in the prediction process, for instance (LR, XG Boost, and RF). The study and models have been applied for one million patients from European Commission (EC), this data set (cough, fever, sore throat, breath, and headache) been considered as a data sensor coming to the proposed system. The aim is to choose the best ML model for Covid-9 prediction. And the results are explained in details with all clarifications and justifications, both models RF and XG Boost have achieved the highest accuracy. For future work, more dataset and additional variables will be investigated with various ML models and compare the performance.

Authors

Amir Nasir Hussein *he is currently a Ph.D. candidate at Al Razi University working on COVID-19 detection using artificial intelligent technologies. He received his B.Sc. from Al-Hussein Eng. College, Iraq and his MSc in networks engineering from Ferdowsi University of Mashhad, Iran. His work as an engineer in satellite news gathering (SNG) at Karbala broadcast channel in Iraq. He can be contacted at amirnasirhussain@gmail.com.*

Seyed Vahab AL-Din Makki *was born in Kermanshah. He received his Ph.D. in Electrical Engineering-Waves from Khaje Nasir Toosi University in 2008. He is with the*

Electrical Engineering Department of Razi University in Kermanshah, since 1994. His current research interests include modern digital radio propagation systems, microwave devices and radio transmitters. He can be contacted at v.makki@razi.ac.ir.

Ali Al-Sabbagh *is a chief engineer at ministry of communication (ITPC-Kerbala) in Iraq. He is also a visitor lecturer at Al-Taff University College (8 years of experience). He got his Ph.D in wireless communication from WiCE Lab at Florida tech, USA since 2019. He obtained his M.Sc from London Brunel University-UK in wireless communication system in 2008. Currently he serves as a technical reviewer in several journals and international conferences. His work and research interests: RF propagation, IoT, and wireless networks. He can be contacted at aalsabbagh@my.fit.edu.*

REFERENCES

[1] Hussein, Amir Nasir, Seyed Vahab Al-Din Makki, and Ali Al-Sabbagh. "Comprehensive study: machine learning approaches for COVID-19 diagnosis." International Journal of Electrical & Computer Engineering (2088-8708) 13.5 (2023)

[2] Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Systems with Applications, 212, 118715. https://doi.org/10.1016/j.eswa.2022.118715

[3] Aslani, S., & Jacob, J. (2023). Utilisation of deep learning for COVID-19 diagnosis. Clinical Radiology, 78(2), 150-157. https://doi.org/10.1016/j.crad.2022.11.006

[4] Hasani, S., & Nasiri, H. (2022). COV-ADSX: An automated detection system using X-ray images, deep learning, and XGBoost for COVID-19. Software Impacts, 11, 100210. https://doi.org/10.1016/j.simpa.2021.100210

[5] Sitharthan, R., & Rajesh, M. (2022). RETRACTED ARTICLE: Application of machine learning (ML) and internet of things (IoT) in healthcare to predict and tackle pandemic situation. Distributed and Parallel Databases, 40(4), 887-887. https://doi.org/10.1007/s10619-021-07358-7

[6] Harshavardhan, A., Bhukya, H., & Prasad, A. K. (2020). Advanced machine learning-based analytics on COVID-19 data using generative adversarial networks. Materials today. Proceedings. doi: 10.1016/j.matpr.2020.10.053

[7] Rahman, M. S., Chowdhury, A. H., & Amrin, M. (2022). Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh. PLOS Global Public Health, 2(5), e0000495. https://doi.org/10.1371/journal.pgph.0000495

[8] Moulaei, K., Shanbehzadeh, M., Mohammadi-Taghiabad, Z., & Kazemi-Arpanahi, H. (2022). Comparing machine learning algorithms for predicting COVID-19 mortality. BMC medical informatics and decision making, 22(1), 1-12.

[9] Chakraborty, C., & Abougreen, A. (2021). Intelligent internet of things and advanced machine learning techniques for covid-19. EAI Endorsed Transactions on Pervasive Health and Technology, 7(26). http://dx.doi.org/10.4108/eai.28-1-2021.168505

[10] Yeşilkanat, C. M. (2020). Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos, Solitons & Fractals, 140, 110210. https://doi.org/10.1016/j.chaos.2020.110210

[11] Zivkovic, M., Bacanin, N., Venkatachalam, K., Nayyar, A., Djordjevic, A., Strumberger, I., & Al-Turjman, F. (2021). COVID-19 cases prediction by using hybrid machine learning and beetle antennae search approach. Sustainable cities and society, 66, 102669. https://doi.org/10.1016/j.scs.2020.102669

[12] Ong, E., Wong, M. U., Huffman, A., & He, Y. (2020). COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. Frontiers in immunology, 11, 1581. https://doi.org/10.3389/fimmu.2020.01581

[13] Wibowo, F. W. (2021, February). Prediction modelling of COVID-19 outbreak in Indonesia using a logistic regression model. In Journal of Physics: Conference Series (Vol. 1803, No. 1, p. 012015). IOP Publishing. DOI 10.1088/1742-6596/1803/1/012015

[14] Prakash, K. B., Imambi, S. S., Ismail, M., Kumar, T. P., & Pawan, Y. N. (2020). Analysis, prediction and evaluation of covid-19 datasets using machine learning algorithms. International Journal, 8(5), 2199-2204.

[15] De Felice, F., & Polimeni, A. (2020). Coronavirus disease (COVID-19): a machine learning bibliometric analysis. in vivo, 34(3 suppl), 1613-1617. DOI: https://doi.org/10.21873/invivo.11951

[16] Tuli, S., Tuli, S., Tuli, R., & Gill, S. S. (2020). Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Internet of things, 11, 100222. https://doi.org/10.1016/j.iot.2020.100222

[17] Cenggoro, T. W., & Pardamean, B. (2023). A systematic literature review of machine learning application in COVID-19 medical image classification. Procedia computer science, 216, 749-756.

[18] Dairi, A., Harrou, F., Zeroual, A., Hittawe, M. M., & Sun, Y. (2021). Comparative study of machine learning methods for COVID-19 transmission forecasting. Journal of Biomedical Informatics, 118, 103791. https://doi.org/10.1016/j.jbi.2021.103791

[19] Zivkovic, M., Bacanin, N., Antonijevic, M., Nikolic, B., Kvascev, G., Marjanovic, M., & Savanovic, N. (2022). Hybrid CNN and XGBoost model tuned by modified arithmetic optimization algorithm for COVID-19 early diagnostics from X-ray images. https://doi.org/10.3390/electronics11223798

[20] Nasiri, H., & Hasani, S. (2022). Automated detection of COVID-19 cases from chest X-ray images using deep neural network and XGBoost. Radiography, 28(3), 732-738. https://doi.org/10.1016/j.radi.2022.03.011

[21] Chumachenko, D., Meniailov, I., Bazilevych, K., Chumachenko, T., & Yakovlev, S. (2022). Investigation of statistical machine learning models for COVID-19 epidemic process simulation: Random forest, K-nearest neighbors, gradient boosting. Computation,

[22] Grekousis, G., Feng, Z., Marakakis, I., Lu, Y., & Wang, R. (2022). Ranking the importance of demographic, socioeconomic, and underlying health factors on US COVID-19 deaths: A geographical random forest approach. Health & Place, 74, 102744. https://doi.org/10.1016/j.healthplace.2022.102744

[23] Jawa, T. M. (2022). Logistic regression analysis for studying the impact of home quarantine on psychological health during COVID-19 in Saudi Arabia. Alexandria Engineering Journal, 61(10), 7995-8005. https://doi.org/10.1016/j.aej.2022.01.047

[24] TRUST. (2021). Zenodo, covid dataset (version 2) available on: https://doi.org/10.5281/zenodo.7316891